THE UNIVERSITY OF ALBERTA

EFFECTS OF DIFFERENTIAL WEIGHTING
ON THE INTER-READER RELIABILITY
OF ESSAY GRADES

by

© JAMES EUGENE CARLSON

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA
AUGUST, 1968

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommended to the Faculty of Graduate Studies for acceptance, a thesis entitled, "Effects of Differential Weighting on the Inter-Reader Reliability of Essay Grades" submitted by James Eugene Carlson in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

The main purpose of this study was to examine the effects of differential weighting on the inter-marker reliability coefficients of essay grades.

Two procedures were used that produce weights for part scores such that the unique contributions of the weighted part scores to the variance of their composite are in a desired ratio. Sets of weights were independently extracted for the part scores of 27 mechanics markers and 21 style and content markers, each of whom marked all of 103 individual essay papers written by grade 12 English students. An analytical grading procedure was used, resulting in 6 mechanics part scores or 16 style-content part scores for each reader on each individual essay paper. Each marker's part scores were weighted and summed to form a weighted composite score for each marker on each paper. The average of the inter-marker correlation coefficients of the weighted composites was then compared to the average of the similar coefficients for the unweighted composites. Although it was found that the resulting composite scores for each of the weighting systems had lower average inter-marker reliability coefficients than had the unweighted composite scores, it was argued that the weighted scores were more valid measures of written-composition skills.

A third procedure, which produced a different set of weights for each marker such that the resulting weighted composites were maximally correlated, was also used. This procedure resulted in weighted composites with higher average inter-marker reliability than that of the unweighted composites but, in the opinion of the writer, the validity of the resulting scores was questionable because of the extremely high weights given some part scores.

Other findings of the study were that the individual markers' weights showed some indication of stability over time and over different essay topics, and that significant proportions of the individual students' essay scores shifted from one gross score classification to another when the scores were weighted.

ACKNOWLEDGMENTS

The writer wishes to express his appreciation to the following persons for their assistance during the development of this thesis.

Recognition is due, in particular to Dr. S. M. Hunka for his constant guidance and encouragement and to Drs. V. R. Nyberg and J. E. Bicknell for their helpful suggestions. Special thanks are due to Dr. Nyberg for permission to use his data for the major portions of this study, and to Dr. Reid and his staff at the Examinations Branch of the Alberta Department of Education for permission to use, and help in collecting the remainder of the data.

The writer is also indebted to Dr. J. Yeager and Mrs. Lois Churchill for constant aid and advice with respect to producing the final copies of the thesis.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

THE PROBLEM

The refinement of measurement instruments is one of the most important tasks presently facing researchers in education, psychology, and related fields. The importance of adequate measurement is perhaps best summed up by the words of Horrocks (1964):

> Whatever the scientific effort, the accuracy of the results and the extent to which the investigator achieves the ultimate objectives of his endeavor depend upon the adequacy and accuracy of the measuring instrument (pp. 55-56).

Criticisms of educational tests and testing procedures have been made by workers within the field of education as well as by others. Campbell and Stanley (1963), for example, pointed out that the adoption of Fisherian methods by educational researchers has resulted in the use of elaborate statistical analysis rather than the improvement of methods of collecting the data to be analysed. On the other hand, Travers (1958) took the stand that:

> Even though the measurement techniques that have been introduced in education are crude, they have permitted a great expansion in our knowledge of the educational process (p. 99).

Although Travers" statement is undoubtedly true, it should not be implied from such statements that educational measurement techniques need no further refinement. The use of

better measurement combined with the use of multivariate statistical analysis could result in even greater expansion in our knowledge of the educational process.

One area of educational measurement that has been the subject of a great deal of controversy is that of essay testing. Nyberg (1966) quoted articles showing that the controversy began towards the end of the nineteenth century, soon after essay tests were first introduced on this continent. The reliability and validity of marks assigned to essays have been questioned and studied intermittently ever since that time. It has been suggested that "in any other science an instrument so imperfect would have been cast on the dust-heap (Ballard, 1925)." but Cast (1939) replied "instead of discarding it, we may seek to improve it (p. 258)." Noyes (Godshalk, Swineford, & Coffman, 1966, introduction), while briefly outlining the history of the experiences of the College Entrance Examination Board with various English composition tests, clearly reported changes that have occurred in the Board's confidence in essay ratings over the last three decades. Perhaps the most important fact that can be learned from such a report is that, despite the often-demonstrated lack of reliability of the marks derived from essay questions, those involved in the measurement of composition skills keep returning to this type of examination question. There seems to be a

prevalent belief that the essay test can measure skills not measurable by objective tests. Anderson (1960) pointed out that this belief has not been empirically supported and expressed the opinion that a valid and reliable means of measuring these skills will not be found "for the simple reason that different examiners disagree about the criteria of excellence or merit to be adopted (p. 95)."

Several examples of measurement specialists at various testing agencies having replaced essay items by objective items were cited by Stalnaker (1951). Criticisms made by these specialists, he pointed out, deserve careful consideration. At the same time, he stated:

> The fact that these specialists in measurement have devised the objective item to overcome some of the weaknesses which they believe to be inherent in the essay question cannot be interpreted as meaning that the essay examination has no useful place in educational measurement (p. 498).

Much expert opinion seems to support the use of essay examinations despite the lack of reliability and validity of the marks assigned to the individual papers. Some of the reasons for the low reliability and validity coefficients have been discussed by Braddock, Lloyd-Jones and Schoer (1963), who mentioned four important variables entering into the rating of compositions: the writer variable, the assignment variable, the rater variable (tendency of a rater to vary in his standards), and the colleague

variable. It is the latter variable, or inter-rater reliability that is the main concern of the present study.

Two methods of marking essays appear to be in general use today: the "wholistic" or "impressionistic" method, in which one mark is given to the essay according to the marker's overall impression, and the "analytic" method in which marks are given separately for various aspects of the content and mechanics of the essay. Several investigators (Coffman & Kurfman, 1966; Coward, 1952) have studied the inter-marker reliabilities of the two methods and concluded that markers can obtain the same degree of reliability with either method.

Cast (1939) mentioned one problem associated with the use of the analytic approach to grading essays:

> the analytic methods, by dealing with numerous isolated and possibly inessential points, may overlook certain general qualities that characterize the essays as a whole (p. 264).

It was then suggested that this difficulty could perhaps be overcome by "some kind of differential weighting (p. 264)." but the suggestion was not clarified. If the marker has completely overlooked some qualities it will be impossible to rectify his oversight by differential weighting. On the other hand, if he has simply overemphasized some qualities at the expense of others, it may be possible to differentially weight the marks he has assigned to the various

qualities. Providing a suitable weighting procedure could be found, it might then be possible to change the emphasis of each marker so that it complied with some predetermined specifications.

One of the findings of Nyberg (1966) was that a mechanics score, made up of six part scores, contributed more to the variance of the composite essay score than did a style-content score made up of sixteen part scores. He suggested that, since the curriculum-makers set a policy of assigning equal marks for mechanics and style-content, the scores on these two marking criteria should contribute equally to the variance of the composite essay scores. In order for the variance contributions to be equal, the two sets of scores would have to be transformed so that they had the same mean and standard deviation before they were added together.

It is conceivable that Nyberg's suggestion could be taken one step further. Since the curriculum makers also designate a maximum mark for each of the twenty-two part scores, it would be desirable to transform these part scores so that they are contributing to the composite mechanics or style-content score in the ratio of the maximum scores set. The problem then, however, is not a matter of a simple transformation in order to equate means and standard deviations. It has been shown (Richardson, 1941, for example) that the contribution of an unweighted variable

depends not only on its variance, but also on its correlations with the other variables. Therefore the intercorrelations of the part scores must be taken into consideration.

The general purpose of the present study was to investigate the effects of differential weighting on the inter-marker reliability coefficients of essay grades assigned through an analytic grading procedure. Specifically, three approaches to the problem of differential weighting of part scores, before summing them to form a composite score, were used. Two of the procedures (Creager & Valentine, 1962; Guilford, 1965) are designed to weight component variables so that the weighted scores will contribute to the variance of their composite in a predetermined ratio. The third procedure (Horst, 1961a, 1961b, 1965) is designed to maximize the sum of the correlation coefficients between weighted linear composite scores. To the writer's knowledge none of these procedures have been applied to part scores assigned by essay readers.

The problems studied were:

1. Would the average of the inter-marker reliability coefficients be increased by any of the weighting systems?

2. Would it be possible to weight the part scores by the Horst procedure and then re-weight by the Creager-Valentine or Guilford procedures, and thus find a set of weights that would result in both a high mean reliability

coefficient, and variance contributions of part scores that contributed to the composite variance in desired ratios?

3. Would the Creager-Valentine and Guilford weights derived for each marker be stable over time and over different topics?

4. Would the Creager-Valentine and Guilford weights be stable over different topics marked at the same time?

5. Would a significant number of essays be affected with respect to gross grade categories by transformations designed to equate the mean assigned scores of all markers?

6. Would a significant number of essays be affected with respect to gross grade categories, when weighted by any of the three weighting schemes?

CHAPTER II

RELATED LITERATURE

Research in Written Composition

Much of the research in the field of measurement of composition skills has been concerned with the effectiveness of various scoring procedures that may be employed in the grading of essays.

Cast (1939, 1940) studied four different grading procedures: the readers' own individual marking methods; marking to see how well the aims of the writer were achieved; marking by general overall impression; and the analytic marking method. It was concluded that the general impression and analytic procedures were the best because they resulted in the most stable marks (average inter-marker reliability coefficient) and the largest first factor (percentage of variance accounted for) when inter-marker correlation coefficients were factor analysed. The average between-marker correlation coefficient, a statistic often used as an estimate of inter-rater reliability, was highest for the analytic method (.492), and second highest for the general impression method (.485). In comparison, the mean coefficients for the individual and the achievement of aims methods were .479 and .439 respectively. It may be noted that the individual method differed little from the analytic

and impressionistic methods on this statistic. Cast reported that discussion with the readers indicated a reason for the similarity; individual methods were essentially individual variations of the analytic and impressionistic methods. Since the same 12 markers were used successively to mark the papers by each of the four methods, the results of this study are open to criticism with respect to order effects. The order was: individual, achievement of aims, general impression, and analytic. There was a tendency for the inter-marker correlation coefficients to increase with successive methods and it may be argued that the order of methods among markers should have been randomized. Cast anticipated such criticisms and answered them by stating that the intervals between marking sessions were long (8 weeks to 2 months), the papers were marked in different orders each time, and the markers reported that they did not remember the marks previously assigned. Since the concern of the study was between-marker reliability, a more important question might be, "did the markers have the oportunity to communicate with each other between marking sessions, and what effect could this have had on the following marks?"

Morrison and Vernon (1941) compared a marking procedure proposed by Steel and Talman (1936) (ST) with a method that was a combination of the impressionistic and

analytic methods (IA). The IA method appears to have been similar to Cast's (1939) individual procedure. In the ST method three columns are ruled off in the margin of the paper. In the first column, minus signs are placed opposite words or idioms that are incorrect, misleading or unintelligible, while plus signs are placed opposite words or idioms that more readily, or more precisely reveal what the author has to say than if he had expressed himself differently. Words or idioms that are merely sufficient and correct receive no mark. Similarly, the second column is used to mark sentence structure, and the third to mark sentence linking. Morrison and Vernon found no difference between the average mark-re-mark reliabilities of the two methods (ST: .860, IA: .859) but mentioned that markers using the ST method expressed unfavorable opinions of the method.

Wiseman (1949) reviewed the results of a marking scheme involving four markers who used an impressionistic procedure, and showed that such marking could result in highly reliable aggregate scores (mark-re-mark r's of .946 and .910 reported). He argued that the four readings took about the same time as one analytic reading and the resulting aggregate mark was more reliable and valid than the results of one analytic reading.

Coward (1952) compared the atomistic (analytic) and wholistic methods of grading on a three hour examination

composed of four essay questions (essay examination in English Composition from the examination for the Foreign Service, 1949). The atomistic markers graded on material, organization, spelling, punctuation, grammar, diction, and rhetoric. The wholistic readers were asked to judge "boldly, decisively, and rapidly (p. 83)" on a ten-point scale. Eight atomistic and eight wholistic readers graded a random sample of 100 papers. Two readers using each method were assigned to each of the four questions comprising the test. The two atomistic readers agreed more closely than did the two wholistic readers for each of the four topics (mean correlation coefficient of .70 as compared to .54), but Coward concluded that the two methods would not have differed in between-marker reliability if the wholistic readers had spent as much time on each paper as was spent by the atomistic readers. This conclusion was based on reliability coefficients estimated, by the Spearman-Brown formula, between two lengthened wholistic readings. Since this formula is based on the effect of increasing the number of scoring units that are summed in calculating the total test score, and equality of the standard deviations of the part scores as well as equality of part score intercorrelations are assumed in its use, Coward's use of it should be interpreted with caution.

In a study of the American History Advanced Placement

Examination, Coffman and Kurfman (1966) reported that

> Analysis of variance procedures revealed significant effects attributable to differences in standards among the four readers, to differences in standards from day to day, and to fluctuations in standards of individual readers from time to time during the reading period (abstract).

Their study was designed so that differences between analytic and wholistic methods of essay reading, with the same readers using both methods, could be examined. Although they stated that "any tendency for the rank order of papers to differ more from reader to reader when the holistic rather than the analytical method is used may be considered negligible (p. 11)." the authors reported no correlation coefficients in support of this statement. The analysis of variance could only show that there was no significant difference between the mean scores of the two methods.

Aside from the studies mentioned above, which deal primarily with differences between different grading methods, most of the studies in the area of essay reading deal with reliability and validity of the scores, and readers. In a comprehensive review, Braddock, Lloyd-Jones, and Schoer (1963) listed 504 studies that have been made in the general area of written composition. Many of these studies were concerned with the reliability of graders, and the conclusion reached by Braddock et al. was that high reader reliabilities are possible. As support for

this conclusion they quoted four studies in which reliability coefficients of from .77 to .96 were reported.

Starch and Elliot (1912) studied the reliability of English grades in general and found a wide range of marks (up to 40 marks out of 100) assigned to the same paper. They used three different groups of readers: 142 English teachers, 86 students in a course on the teaching of English, and 98 students in an educational measurement course (mainly superintendants and principals). The ranges of marks for the three groups of markers were about the same, but the students and the teachers who were from small schools tended to grade more leniently than did teachers from large schools and the superintendants and principals.

An average correlation coefficient of .738 was found between six markers by Finlayson (1951). He stated that marker consistency should not be the only criterion used to estimate the reliability of essay grades and proceeded with several other analyses. A centroid factor analysis of the matrix of correlations between markers yielded only one factor, with marker loadings ranging from .779 to .901. Also reported were: average mark-re-mark correlation coefficients of .937 and .941 for the sums of four markers' grades on two different essay topics (individual's mark-re-mark r's ranged from .731 to .966); an average reliability coefficient of .691 between the grades assigned to two

papers written by the same students one week apart; and a reliability coefficient of .786 between different readers and different topics. In the latter case, two different sets of markers (three markers in each set) read the two sets of papers, and thus the reliability coefficient was being estimated between different topics and different readers. In an attempt to establish whether or not the scores had validity, Finlayson calculated the correlation coefficient between essay grades and teachers' estimates of the pupils' performance in English (.776) and the Moray House English test (.755). An analysis of variance indicated that the performance of children varied significantly from one essay to another and Finlayson concluded that the performance of a child on one essay topic is not a reliable estimate of his ability in written composition. There were also significant differences between markers.

Torgerson and Green (1952) in a factor analytic study of 38 individual essay papers found that the first centroid factor extracted from a matrix of correlation coefficients between 20 readers accounted for 67 percent of the total variance but that there were three other factors present. The papers were marked in two parts so that a style score made up of five part scores and a material and organization score made up of three part scores were found. The markers who had the highest loadings on the second factor had the

highest correlations between their style and material scores. Aside from sex differences in factors two and three, no other interpretations were found possible. Torgerson and Green were interested in studying the dimensionality of the correlation matrix and in evaluating factor analysis as a procedure for constructing and evaluating achievement tests. They concluded that the strong first factor indicated a considerable agreement among readers and that the factor analytic procedure should be of great value.

Two studies were reported by Huddleston (1954). The object of the first was to determine whether ability to write, as defined by ratings of writing ability, could be more reliably measured by objective English tests than it could be measured by essay examinations. The variables used included: an objective English test with 27 items designed to measure punctuation, 33 to measure idiomatic expressions, 47 to measure grammar, and 70 to measure sentence structure; three twenty-minute essay questions; a paragraph revision test; a verbal test made up of antonyms from the College Entrance Examination Board Scholastic Aptitude Test (SAT); English course grades; and ratings of English instructors. The subjects were 413 freshmen from 16 different English classes at 5 different colleges. In order to estimate the reader reliabilities of the essay and paragraph revision tests, a random sample of

papers was selected and each paper read by two different readers. The correlation coefficients between readers were found to be .78 for the essays (scores made up of the sums of the marks from the three essays), and .83 and .59 for the two paragraph revision questions. It should be noted that the sample for the second paragraph revision test was rather small (38 papers), while those for the first paragraph revision and the essay questions were larger (89 and 122). The smaller sample could possibly have resulted in a restriction of the range of the marks and thus a smaller correlation coefficient. Kuder-Richardson formula 21 coefficients were used as estimates of the reliabilities of the objective English and the verbal tests, and these coefficients were found to be .93 and .85 respectively. Each essay paper was rated for: material and organization; spelling; punctuation; syntax; vocabulary; and sentence structure. The correlation coefficients between the total scores of the three essay questions were found to be .41, .41, and .32. Using the Spearman-Brown prophecy formula it was found that an essay test composed of three equivalent twenty-minute essay questions, correlating .41 with each other, would have a reliability of .68. Writing three equivalent essay items would, of course, be an extremely difficult task. The essay test (3 essays combined) was found to be the best predictor of course grades (median of

r's for the 16 classes = .43), followed by the verbal test (median r = .38), objective English test (.34), and paragraph revision (.23). For predicting instructors' ratings, however, the best predictors were the objective English and verbal tests (median r's both = .43) followed by the essay test (.34) and paragraph revision (.27). Huddleston concluded that the essay test had a language ability component as well as a verbal component.

In the second study, Huddleston used a total of 763 high school students (due to missing data N's varied from variable to variable), and the following variables: objective English test--45 of the items from the original test, chosen by item analysis of the first study results; essay test--one of the essays from the original study; paragraph revision test--same as in original study; scores on the verbal sections of the SAT; high school instructors' ratings; and high school English course grades. The reliability coefficients for the objective English Test and SAT verbal were .78 and .96 respectively (odd-even correlations corrected by Spearman-Brown formula). In this study, 15 readers read each of the paragraph revision questions, and 39 read each of the essay papers. The "approximate true score" of a paper was defined as the average score assigned by readers, and individual readers were judged to be reliable to the extent that their scores were similar to the

true scores. A method devised by Tucker (Huddleston, 1954, pp. 191-192), was used to estimate reliability coefficients. The square of the correlation coefficient between an individual reader's scores and the true scores was used as an estimate of the reader's reliability. For the essay, the average reader reliability was found to be .62, as compared to .81 and .84 for one paragraph revision question (2 different sets of markers), and .69 and .72 for the second paragraph revision question. The paragraph revision test had low correlation coefficients with other variables and Huddleston concluded that it was uncertain what was being measured by the test. The verbal test was found to be more closely related to writing ability than were any other variables, correlating .76 and .77 respectively with the two criterion measures: instructors' ratings, and average English grades. For the instructors'-ratings variable all instructors made independent ratings of their students on two different occasions and the reliability coefficients as estimated by Spearman's rhos ranged from .86 to .999. The two criterion variables were highly related (r = .82). A method of multiple-regression adjustment devised by Tucker (Huddleston, 1954, p. 194) was used in order to equate criterion-variable scores of subjects from different English classes. Multiple linear regression analysis was used in order to study the components of writing ability and it was

found that the other variables failed to add appreciably to the proportion of writing-ability variance predictable from the verbal test alone. The zero order correlation coefficients between verbal test scores and the adjusted scores on the two criterion measures were found to be .76 (instructors' ratings) and .77 (average English grades), while the coefficients of multiple correlation with all predictors were .79 and .80 respectively. In the second study, unlike the first, it was found that essay scores were more highly correlated with verbal than with objective English scores. Huddleston concluded that ability to write was nothing more than verbal ability and that the best estimate of such ability is the verbal score of the Scholastic Aptitude Test.

A series of articles (Peel & Armstrong, 1956; Penfold, 1956; Pidgeon & Yates, 1957; Wiseman, 1956) were written on the use of essays for eleven-plus selection in England.

In the first article, Penfold (1956) described a study in which 15 readers each graded all the essay papers written by 165 candidates. An analysis of variance technique indicated that markers' means differed significantly. Some errors of interpretation of this data have been pointed out by Lawley and Pilliner (1957).

Peel and Armstrong (1956) in the second article of the series reported a study that was largely correlational in

nature. They showed that a criterion variable of Grammar-school teachers' assessments of performance was more highly related to an objective English test (Moray House: average of the r's of 11 schools = .47), than to an English composition (essay) test (average r = .44) or an intelligence test (Moray House: average r = .40). Using a combination of the teachers' assessments of performance and grades in a modern language as a criterion variable, it was found that the best combination of two predictors was the essay plus the intelligence test for six of the eleven schools involved. The authors reported that in eight out of the eleven schools there was a fall of .03 or less in "maximum prediction" when the objective English was eliminated from the prediction. Maximum prediction appears to have been defined as a multiple correlation coefficient rather than as its square which is usually used as an index of proportion of predictable variance.

The third study of the series was that of Wiseman (1956), who supported Finlayson's (1951) contention that marker consistency should not be the sole criterion of essay reliability and, in order to examine test-retest reliability, had each student write two papers. The same four readers graded both sets of papers and the correlation coefficients between the two sets were found to be .772, .803, .796, and .816 for the four markers. A correlation

coefficient of .894 between aggregate marks on the two sets (presumably the sum of the four grades assigned to each paper) was also reported. Since it was found that the sample had a restricted range on the Moray House English Test (low standard deviation compared to that of the standardization sample) it was assumed that the same restriction would occur in the essay scores. A correction for restriction of range was therefore applied to the aggregate mark correlation coefficient, yielding a corrected coefficient of .923 which was considered to be adequate as an estimate of test-retest reliability. This use of the correction is, however, a questionable procedure since it was assumed that the standard deviation of the essay marks should be the same as that of the Moray House exam. Since the sources of variation in essay ratings are likely to be very different from those of an objective examination, this assumption is likely to be unwarranted. Wiseman also attempted to estimate the validity of the tests but admitted that his criterion variables (school results, estimate of overall school attainment, and estimate in written English) were unsuitable.

Pidgeon and Yates (1957), in the final study of the series, had 254 students write essays on three different topics on three different occasions, and then had them write an additional paper on the first topic. The seven readers used in the study each marked papers on each topic

and also re-marked the original papers that they had marked on the first topic. Two readers graded each paper. One of the findings of this study was that there were two groupings of markers with respect to mean assigned scores--a high group and a low group. There were significant differences between the means of markers from different groups but no significant differences between means of markers within each group. The upper group consisted of the two female markers and the lower group of the five male markers. Pidgeon and Yates also found that there was a practice effect on the essay papers, that is, that mean grades increased for each subsequently written set of papers. That this increase was a function of the students' papers, and not of a tendency for readers to grade more leniently, was indicated by the fact that in re-marking the first papers, after marking papers on other topics, readers gave lower mean scores than they had previously given to the same papers. Analysis of variance procedures were used to estimate the reliability coefficients, shown in Table 1, for the essays and for a more objective English test. The items of the latter "were of the creative response type (p. 38)" but "Some of them admitted to such a restricted range of correct answers as to allow a nearly objective form of scoring (p. 38)." The authors also examined the validity of essay and objective English tests using as a

criterion the order-of-merit estimates made by the heads of eight schools on a total of 437 students. The correlation coefficient between the criterion and the objective English test was .724 as compared to correlations of .578 and .549 for two essay topics.

Table 1

Reliability Coefficients Reported by Pidgeon & Yates

Type of Reliability	Essay Question	Objective Test
Different examiners marking same papers	.815	.979
Same examiners re-marking same papers	.866	.985
Same examiners marking same test questions written a second time	.766	.850
Different examiners marking same test questions written a second time	.719	.845

Remondino (1959) like Torgerson and Green (1952), used factor analysis in an attempt to study the manner in which readers discriminate between essays. Unfortunately, the article does not clearly present the procedures used for some parts of the analysis. This lack of clarity may result from the fact that the article cited herein is a

translation of the original[1]. A total of 17 variables, identified by teachers as useful in the grading of essays, were identified. A team of four teachers and six non-teachers (of "average educational ability") read 230 compositions written by 15 and 16 year old boys attending courses in a vocational school. Each paper was rated for each of the 17 variables on a three point scale. It would appear that, for each marker, a matrix of correlation coefficients between pairs of the 17 variables was then analysed for clusters of the variables, by the method of B coefficients (Holzinger & Harman, 1941, pp. 23-34). The data for this analysis were based upon "about one-third of the essays chosen at random (p. 244)." The clusters were quite similar for all markers and there appeared to be no systematic differences between teachers and non-teachers. Remondino then asked three of the four teachers to each read all 230 papers, calculated matrices of intercorrelations between the 17 part scores for each marker, and found "a single average matrix (p. 246)." which was factor analysed by the centroid method with an oblique rotation. This analysis resulted in four factors which were very similar to the clusters found in the initial stages of the study. The four factors were labeled: graphic representation, language

[1] Paper presented at the International Congress of Applied Psychology, Rome, April, 1958.

usage, content and arrangement of the essay, and personal aspects (maturity, originality, imagination, etc.) of the essay.

Anderson (1960) used an analysis of variance technique to examine differences between Testing Occasions (TO), Markers (M), Tests (T), and Marking Occasions (MO). Subjects were 55 grade eight students and they were administered two forms (A and B of Level 3) of the Cooperative Test Division's STEP Essay Test, one in the morning, and one in the afternoon of the same day. On three other testing occasions, at weekly intervals, subjects wrote essays on the same two topics so that each wrote four compositions on each form, on four different testing occasions. Three readers "well practiced in using the marking schedule (p. 97)" graded the essays on a seven-point scale. The analysis of variance showed significant differences (.01 level) for TO, M, and T main effects, and for the M x MO interaction. No other differences were significant at the .05 level. Anderson concluded that the STEP Essay Test had not reduced into insignificance the variability attributable to "well-known sources in the marking of essays--composition fluctuation, the unrepresentativeness of essay-samples, and discrepancies among markers (p. 101)."

Validity of essay ratings was the main concern of Young (1962). Teachers in three schools were asked to

assign three essays to their students, and to have each paper marked by three members of the school faculty, so that nine judgments of ability in written composition resulted for each student. The essays written as part of the eleven-plus classification examinations were then used in order to equate the marks assigned at different schools. The eleven-plus papers were marked by placing them in order-of-merit, using Wiseman's (1949) impression method with a slight change of emphasis. Since several markers rated these papers, each marker's ranking was fitted to the normal curve, then rescaled to scores having mean 100 and standard deviation 15. Each teacher's scores (sums of the nine judgments) were then scaled in accordance with the distribution of scores of his pupils on the classification essay. Using the grades that 121 students achieved the following year in grammar school as a criterion variable it was found that the validity coefficient of the teachers' ratings was of the same order (.70) as marks from standardized arithmetic (.71) and English tests (.74), lower than that of a verbal reasoning test (Moray House) (.87), and higher than that of the eleven-plus classification essay (.42). The validity coefficient based upon a sum of verbal reasoning, arithmetic, English, and the teachers' essay marks was found to be .93. Once again the criterion variable was the grammar school grades.

In order to study the validity of marks assigned to essays by readers using the general impression marking technique, Britton (1963) collected five pieces of written composition from each member of a sample, representative by grade, of eleven plus candidates (168 subjects drawn from 692 candidates). Each piece was independently graded by two examiners and the resulting marks were used as a pooled assessment of the student's ability level in written composition. Eight other examiners then marked the eleven-plus compositions by rapid general impression on a 20 point scale. In addition, one of the eight markers was subsequently asked to give each paper a mark for mechanical accuracy based upon about the first 300 words. Each eleven-plus paper was also marked by one of the two official eleven-plus readers using an analytic procedure. High validity coefficients (r's with pooled assessment) were found for the impressionistic graders (.76 for a randomly selected group of three markers; .81 for another set of three markers' scores, with the mechanics scores added to them). The coefficient found for the analytic markers was .71. Consistency of marker ratings as shown by mark-remark correlation coefficients ranged from .74 to .95.

Diederich (1966) reported a study[2] in which 300 papers

[2] Diederich, P. B., French, J. W., & Carlton, S. T., _Factors in judgments of writing ability_. Research Bulletin 61-15, (out of print). Princeton: Educational Testing Service, 1961.

were graded by 60 readers from six different fields. They were asked to sort the papers into nine piles in order of general merit. None of the 300 papers received fewer than five different grades and 101 received all nine grades. The average correlation coefficient between the 60 readers was .31 and between the English teachers was .41. The matrix of between-reader correlation coefficients was factor analysed and the conclusion reached from the results was that the readers were responding to different qualities in the papers. Since five factors were found and defined, it was suggested that readers should be asked to grade papers on the five defined components. When this marking scheme was later tried out in three large high schools, factor analysis revealed only two factors, which were labeled "general merit" and "mechanics". No reliability coefficients were reported for this data. The present writer feels that reliability should have been higher since markers were given specific criteria on which to rate papers.

Nyberg (1966), in a factor analytic study of marker ratings, used 103 essays written by twelfth-grade Alberta students as part of the Alberta Department of Education, High School and University Matriculation Examinations Board, Departmental Examinations. Twenty-seven readers graded the essays on six mechanics variables and twenty-one different readers graded the essays on sixteen style and content

variables. Principal axis factor analysis of the intercorrelations of style-content readers, with quartimax rotation, revealed three factors. Sixty-two percent of the total variance was accounted for by the first factor, on which all readers had loadings of at least .64. The reliability of the style-content readers (average between-marker correlation coefficient) was found to be .595. Those readers having loadings of at least .80 on factor I and whose mean scores fell between 64.01 and 70.73 (within 3 standard error of measurement units of the overall mean of 67.4) were judged to be suitable readers and the average correlation coefficient between these readers (9 in all) was .73. A similar analysis of the mechanics markers' scores revealed a first factor accounting for 77.8 percent of the total variance and all first factor loadings were .80 or greater. The average between-marker correlation coefficient was found to be .768, but the markers' mean assigned scores varied a great deal (from 40.0 to 72.4 on a 175 point scale). The mean correlation coefficient between the 19 markers classified as suitable was found to be .81. Nyberg also found that, when scores were totaled, the mechanics scores contributed much more to the composite variance than did the style-content variables.

In the second part of his study, Nyberg factor analysed a matrix of correlation coefficients between essays'

true scores, in order to study the papers themselves. True scores were defined by taking the average of all markers' scores on each of the marking variables, and then summing these average part scores for each paper. Three resulting factors were interpreted. Papers with high loadings on the first factor were those with a high level of proficiency in the higher aspects of mechanics (word usage, sentence errors and grammar). Papers with high loadings on the second factor showed agreement (high r's) between the mechanics and style-content part scores, while those loading on the third factor showed lack of agreement between these two groups of part scores.

In a third factor analysis, this time of the true scores of the marking variables, Nyberg found that variables used to rate general impression in style and content loaded highest on the first factor, higher mechanical skills variables loaded highest on the second, variables related to the plan of the essay loaded highest on the third, and four other factors were defined by the one variable loading on each: form, handwriting, punctuation and spelling.

Godshalk, Swineford and Coffman (1966) studied the reliability and validity of various types of questions designed to measure writing ability.

The criterion measure for the validity study consisted

of five samples from each of 646 grade 11 and 12 students' written composition (5 common topics), graded by 25 experienced readers who used a wholistic method. The students "represented both public and independent schools of various sizes in cities and smaller communities, and in all commonly listed major geographical areas of the country (p. 9)." The samples of writing were each scored independently by five different readers on a three point scale, papers being distributed to readers such that all readers graded at least one paper produced by each student. These scores were then summed (thus the score for each paper could range from 5 to 15, and for each student from 25 to 75) and score and reading reliability coefficients of the resulting scores were estimated from the results of a three-factor analysis of variance (Students x Topics x Readings). Based upon the total scores, an estimate of reliability of readings, of .921 was found by

$$\frac{\text{Students Mean Square} - \text{Error Mean Square}}{\text{Students Mean Square}}$$

and an estimate of reliability of scores of .841 by

$$\frac{\text{Students Mean Square} - \text{Student x Topics Mean Square}}{\text{Students Mean Square}}$$

Single topic reliability coefficients based upon five readings, as estimated by analysis of variance ranged from .739 to .777. The analysis of variance also revealed a significant Topic effect indicating that the assigned grades

varied significantly from topic to topic, and a significant Reading effect indicating that grades varied significantly from reading session to reading session. Examination of reading session means indicated that the latter effect resulted from a "more generous" grading in the first session. A significant Student x Topic interaction effect was also found, indicating that different students did better on different topics.

In order to further examine the reliability of the criterion essays, Godshalk et al. calculated average correlation coefficients between single readings within each topic (.361 to .411), and between topics (.221 to .308), as well as correlation coefficients between total scores on the five readings of each topic (.435 to .592).

As a result of this study of a written-composition criterion measure consisting of the sum of scores from five samples of writing, Godshalk et al. recommended the use of multiple readings, rather than the use of methods designed to increase the reliability of single readings.

After having established the reliability of the criterion, Godshalk et al. used the following as measures of writing ability and tried to validate them against the criterion.

1. paragraph organization questions (scrambled sentences, to be rearranged)

2. objective test of English usage

3. sentence correction exercises

4. prose groups (paragraphs with a sentence missing and a choice of four sentences to add where the sentence is missing)

5. error recognition (indicate which of four classes of errors, if any, occurs in a sentence)

6. construction shift (decide additional changes necessary in a sentence when a specified change is made)

7. & 8. two interlinear exercises (re-write poorly written material)

9. Scholastic Aptitude Test (SAT) and Preliminary Scholastic Aptitude Test (PSAT)

The six objective English tests (variables 1 through 6) had correlation coefficients of .458 to .707 with the criterion and the two interlinear exercises had coefficients of .644 to .668. Godshalk et al. also calculated validity coefficients for various combinations of the variables taken three at a time, and these ranged from .706 (variables 1, 4, and 8) to .767 (variables 2, 6, and 8). They concluded that objective questions designed to measure writing ability were valid when evaluated against a reliable criterion. They attributed the difference between their findings and those of Huddleston (1954) to the use of a more reliable criterion variable in their study.

In order to estimate the validity of a twenty-minute essay as one item on a test of written composition, Godshalk et al. tried using one of the criterion essays as a test variable, thus reducing the criterion measure to the sum of marks on four essays. Different combinations of, either one of the interlinear exercises or one of the essays, with some of the objective tests plus the verbal score from the PSAT or SAT were formed. It was found that, in all cases, a combination including an essay resulted in a higher multiple correlation coefficient with the four-essay criterion than did the same combination with an interlinear exercise replacing the essay (.733 to .796 as compared to .712 to .757).

One year later, Godshalk et al. carried out an extensive field trial of the reading of a sample of 533 of the original papers (individuals for whom PSAT scores were available) by 145 readers. Each paper was read either four or five times and some readers used a four point scale rather than the original three point scale (there had been some indication in the previous study that, when in doubt, readers tended to use the middle category). In general, the reliability coefficients found for the field trial were higher than those for the original readings, and those for the four point scale were higher than those for the three point scale. The higher validity coefficients on the field

trial were attributed to the larger number of readings of the essays. Correlations between scores based upon five original readings of an essay and the four-essay criterion ranged from .508 to .658, while those between scores based upon four field trial readings ranged from .542 to .640, an average of .011 higher. Thus the authors concluded that there was no loss in validity in moving from the experimental to the field trial readings.

The main conclusions reached by Godshalk et al. were:

> 1. The reliability of essay scores is primarily a function of the number of different essays and the number of different readings included (p. 39).
>
> 2. When objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove to be highly valid (p. 40).
>
> 3. The most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions or interlinear exercises in combination with objective questions (p. 41).

To find out whether or not reader reliabilities would be as high as those found by Godshalk et al. when 80,000 papers were being read over a five day period by 145 readers, was the purpose of a study by Myers, McConville and Coffman (1966). The essays were read wholistically and assigned grades of from one to four. A sample of twenty-five papers was selected and each day 25 of the markers were randomly selected to rate this sample of papers along with the other essays they were to read. Readers did not know

which papers were to be used in the study. An analysis of variance procedure was used to estimate the average correlation coefficient between judges by comparing the variance between papers with the variance within papers. The Spearman-Brown formula was used in order to step down the resulting coefficients and thereby estimate the reliability of a single judgment. The resulting coefficients for the five days were .466, .364, .493, .476, and .264 respectively. When these coefficients were stepped up in order to estimate the reliability of four readers, the coefficients ranged from .589 to .795. In order to see if the markers had a small number of points of view with respect to grading, the matrix of covariances between papers was factor analysed and a varimax rotation used. Four factors were extracted and it was found that the poor papers tended to load on one factor while the good papers loaded on a second factor. No attempt was made to define the four factors but the authors argued that the factors represented different points of view in the relating of grades to the quality of the papers. It should be pointed out that in using the Spearman-Brown formula, the researcher assumes that the additional marking units are equivalent to the original units. This assumption may be difficult to justify in the case of additional readings of essay papers.

A study by Bracht and Hopkins (1968) raises some

questions about the assumption that written compositions measure skills not measurable by objective tests. The achievement of 279 college sophomores in a course in Educational Psychology was evaluated, at the end of each of two units of the course, by essay and objective examinations, allowing equal testing times for each type of test. The four instructors in the course graded the papers in their usual manner and subsequently participated in an extensive evaluation of 100 randomly selected papers. The objective tests consisted of multiple choice items. Kuder-Richardson formula 20 reliability coefficients of .40 and .50 were found for the two objective tests (24 items each) and Spearman-Brown estimates indicated that 60 item tests would have reliabilities of .63 and .71. An analysis of variance approach was used to estimate the reliability coefficients of the original essay readings. The coefficients were found to be .58, .16, and .69 (for the second unit two different sets of essay questions were constructed so that examinees writing the exams early could not inform later examinees about the questions). The investigators found that, in general, correlation coefficients between essays and many other variables (sex, age, SAT, STEP Writing, penmanship, appearance of essays, English grades, etc.) were of the same magnitude as those between objective tests and the other variables. It was found that grades assigned by

some readers were consistently related to factors not relevant to academic achievement (penmanship, order of reading). Thus the data of Bracht and Hopkins do not support the assumption that essay and objective tests measure different abilities. The authors pointed out however that their findings cannot be validly used to generalize to different kinds of subjects, courses and topics.

Page (1966) carried out a study of a rather different nature than those reported above. He used a computer to grade 138 essay papers written by high school students in grades 8 to 12. The papers were also graded by four English teachers and between-marker reliability coefficients involving grades assigned by the computer program were indistinguishable from those involving only the English teachers. The overall range of correlation coefficients was from .44 to .61. Page optimistically predicted that in the future the computer will be able to grade more reliably than can human judges. One cannot help wondering what the outcome of Page's experiment would have been had his teachers been more reliable graders. Many of the studies cited above resulted in higher inter-marker reliability coefficients than those reported by Page.

In summary, various investigations have shown that reader reliability as estimated by the average correlation coefficient between readers, ranges from .31 to .738,

mark-re-mark correlation coefficients are generally found to be in the .70 to .90 range, and estimations of reliability of several readers over several essay topics can be as high as .94 when estimated by analysis of variance methods.

It is interesting to note that very few of the studies described above make any mention of differences in means of marks assigned by different readers. If a number of different readers are grading essays, and some readers have a tendency to mark lower (have lower mean assigned scores) than others, those papers read by such markers may be unjustly assigned lower scores than other papers. Since the correlation coefficient between two sets of scores is independent of the means of the scores, such differences in means may occur even when there is high marker reliability. Nyberg (1966) considered this problem when he chose as "suitable" those readers whose mean assigned score fell within three standard error of measurement units of the overall mean. In several studies (Coffman & Kurfman, 1966; Penfold, 1956; Pidgeon & Yates, 1957) significant differences between markers' mean assigned scores have been found. Young (1962) transformed his classification essay scores in order to equate markers' means.

Another problem not discussed in most articles dealing with essay reliability is the effects of differences between readers' standard deviations. Such differences,

unlike differences between means, could affect the values of the reliability coefficients if the range of scores should become sufficiently attenuated for some markers.

The widespread use of the technique of averaging correlation coefficients also deserves some comments. It has been shown that the sampling distribution of the correlation coefficient becomes skewed as the parameter estimated by r departs from zero. Because of the non-normal distribution a transformation developed by Fisher (described by Guilford, 1965, pp. 348-349) should be used prior to averaging. Two studies in which this procedure was used were those of Pidgeon and Yates (1957) and Young (1962) cited above.

Weighting of Part Scores

Few articles dealing with the weighting of part scores from essay examinations can be found in the literature. Stalnaker (1938) discussed such weighting but most of the examples cited to support his views were based upon objective test scores rather than on essay scores. It was concluded that

> In the ordinary subject-matter achievement examination which deals with a single field, and in which a number of part scores are assigned, the simplest and wisest rule for the readers to follow is to assign the maximum values for the parts according to the number of degrees of difference in the answers which can be consistently judged. The total score may then be obtained by merely adding the assigned

> part scores. The influence of the usual weighting factors is so small as to be insignificant (1938, p. 490).

but at one point it was mentioned that

> In essay tests where the number of scorable units (items) is smaller, or where sections, rather than items, are to be combined, weighting may have a greater influence (p. 486).

Although Stalnaker did not discuss weighting in a rigorous mathematical manner, he referred to discussions with Wilks, who did (1938) discuss mathematical considerations of the weighting of part scores. Wilks first showed that, for a large number (n) of correlated part scores, two different linear functions of the same scores will be highly correlated. It was shown that the correlation coefficient between such pairs of functions differs from unity by terms of order 1/n . The implication is that, if one is dealing with a large number of part scores, the use of different weighting systems will have little effect on the ordering of individual scores. Wilks then discussed three methods of weighting part scores before combining them.

In the first method, weights are derived so as to minimize the generalized variance of weighted part scores of all individuals receiving a given composite score. Generalized variance was defined as follows.

> In the case of two variables, if we form triangles with all possible pairs of individuals (each represented by a point) using the means of the two variables as coordinates of the third point in each

> case, square the area of each triangle and average the squared areas for all pairs of individuals in the population, the result multiplied by $(2!)^2$ is the generalized variance of the bivariate distrubution. In the case of N dimensions, N individuals and the point representing the means of the N variables determine an N-dimensional simplex, i.e., an N-dimensional generalized tetrahedron. If we form all possible such simplexes, square their volumes, and take the mean value of the squared volumes, multiplying by $(n!)^2$ we get the generalized variance of the N-variables (1938, p. 30).

Wilks pointed out that Horst (1936) and Edgerton and Kolbe (1936), although they used different criteria for their derivations, arrived at weights that were proportional to those of the minimum generalized variance method. Horst's criterion was the maximizing of the variance of the weighted composites with the restriction that the sum of the squared weights equalled unity: Edgerton and Kolbe's criterion was the minimizing of the sum of the within-person (over part scores) variances, while keeping the sum of the squared weights constant. Horst (1941, pp. 72-73) made two observations about the weights derived by this system: the weights obtained depend upon the units of measurement of the original variables; and the weights so derived are proportional to first principal axis factor loadings (also mentioned by Guttman, 1941, p. 346). The first observation may well apply to any of the weighting schemes discussed herein. Wilks (1938, p. 38) also made an observation about these weights that makes them appear rather undesirable for

weighting part scores of essays. If a variable is independent of (uncorrelated with) all other variables in the set, it will receive a weight of zero. It seems, to the present writer, hard to justify the exclusion of a part score from the composite just because it is independent of other part scores, and in fact such variables may be very important to the overall composite mark representing a broad concept of skills in written composition.

The second procedure discussed by Wilks (1938) involves weighting in order to equalize the correlation coefficients between each sub-test and the total score. The possibility of weighting such that the correlation coefficients between part scores and their composite are in a predetermined ratio rather than being equalized was also discussed. In an article dealing with the combination of measures, Richardson (1941) also discussed this kind of weighting. It was shown that the composite variance could be found by the use of the relationship

$$\sigma_c^2 = \sigma_1 \sum_{i=1}^{n} (r_{1i}\sigma_i) + \sigma_2 \sum_{i=1}^{n} (r_{2i}\sigma_i) + \ . \ . \ . + \sigma_n \sum_{i=1}^{n} (r_{ni}\sigma_i)$$

(equation 13, p. 383)

where: σ_c^2 is the variance of the composite of n variables

r_{ji} is the correlation between variables j and i

σ_i is the standard deviation of variable i

Richardson then defined the contribution of a variable (j for example) to the composite variance as the term of the form

$$\sigma_j \sum_{i=1}^{n} (r_{ji}\sigma_i)$$

in the right hand side of the preceeding equation. This definition of variance contribution will be further discussed and compared to other procedures in a subsequent section of this paper.

Wilks (1938) thirdly discussed weighting in order to equalize the variance contributions of part scores, or alternatively in order that the part scores contribute to the total variance in a given ratio. Creager and Valentine (1962) proposed such a procedure and defined the unique contribution of a part score as

> the difference between the squared multiple from the entire set of components (1.00) and that from a subset with the given component, or components, missing (p. 33).

This difference may be expressed as

$$R_c^2 - R_{c,n-i}^2 \qquad (1)$$

where: R_c^2 is the squared multiple correlation coefficient between the sum of all part scores and their composite (and therefore = 1.0)

$R_{c,n-i}^2$ is the squared multiple correlation coefficient between the sum of all part scores

(composite) and the sum of all part scores except i

Equation (1) is therefore equivalent to

$$1.0 - R^2_{c,n-i} \qquad (2)$$

This quantity is defined as the proportional unique contribution (U_i) of part score i to the composite variance. The actual variance contribution (in raw score units) of a part score may thus be found by

$$\sigma^2_c \; U_i = (1.0 - R^2_{c,n-i}) \; \sigma^2_c \qquad (3)$$

Creager and Valentine then showed how a set of weights can be found that will bring the unique variance contributions of several part scores into a desired ratio. The square of the desired weight for a variable when multiplied by the unique variance contribution produces a product that is a factor of the desired contribution of that variable. For example, if the unique variance contributions of three variables ($\sigma^2_c \; U_i$) are 33.6, 21.9, and 5.0, and it is desired that the variables contribute 10, 25, and 10 percent of the variance respectively, the weights are found by[3]:

$$\begin{aligned} W^2_1 \times 33.6 &= 100 \\ W^2_2 \times 21.9 &= 250 \qquad (4) \\ W^2_3 \times 5.0 &= 100 \end{aligned}$$

[3] Example from Creager and Valentine (1962), figures rounded.

In actual fact, 10, 25, and 10, or any numbers in the same ratio could be used on the right hand sides of these equations. Although this would change the values of the weights, the ratio of weights would be the same, and the contributions of the weighted variables would be in the same ratio for all such sets of weights. Similarly there is no need to multiply the U_i by σ_c^2 before solving the equations in (4). In this example, Creager and Valentine were able to find weights that, when applied to the three variables, resulted in weighted variables having percentage contributions of 10.6, 26.6, and 10.6 when the percentage contributions of the unweighted variables had been 30.7, 20.0, and 4.6 respectively. It can be seen that the contributions of the weighted variables are very close to the desired values.

In general, the following relationships were employed by Creager and Valentine (although they did not express them in this manner):

$$W_1^2U_1\sigma_c^2 : W_2^2U_2\sigma_c^2 : \ldots : W_i^2U_i\sigma_c^2 : \ldots : W_n^2U_n\sigma_c^2 \tag{5}$$
$$= D_1 : D_2 : \ldots : D_i : \ldots : D_n$$

and therefore

$$W_i^2U_i\sigma_c^2 = kD_i \tag{6}$$

where k is a constant, which may be absorbed into D_i, thus

$$W_i^2 U_i \sigma_c^2 = d_i \tag{7}$$

and thus the quantities on the right hand sides of the expressions in (4) are any d_i such that

$$d_1 : d_2 : . . . . : d_i : . . . . : d_n \tag{8}$$

expresses the desired ratio of variance contributions. It might also be pointed out again that the σ_c^2 term is a constant and may therefore be dropped from expressions (5) through (7).

Hazewinkel (1963) showed that the following relationships hold:

$$U_i = 1.0 - R_{c,n-i}^2 \tag{9}$$

$$= \frac{\sigma_i^2}{\sigma_c^2} (1.0 - R_{i,n-i}^2) \tag{10}$$

$$= \frac{\sigma_i^2}{\sigma_c^2 R^{ii}} \tag{11}$$

where: $R_{i,n-i}^2$ is the squared multiple correlation coefficient between variable i and the n-1 other variables

R^{ii} is the ith diagonal element of the inverse of the n by n predictor intercorrelation matrix

The unique contribution of a variable to the composite

score variance has been defined differently by Chase (1960). He defined the unique contribution of a variable (i for example) as the square of the least squares weight of that variable (β_i^2) when all the variables are placed in a regression equation to predict the composite. That is, if

$$Z_c = \beta_1 Z_1 + \beta_2 Z_2 + . . . + \beta_n Z_n \qquad (12)$$

where: $Z_1 . . . Z_n$ (N by 1) are column vectors of standardized scores of each of N subjects on each of n component variables

Z_c (N by 1) is a column vector of unweighted composite scores ($Z_1 + Z_2 + . . . + Z_n$) for each subject

$\beta_1 . . . \beta_n$ are least squares weights

is solved for the weights that will maximize the correlation between the composite and the weighted composite, the squares of the weights are defined as the proportional unique contributions of each variable to the composite.

Guilford (1965, pp. 423-426), using Chase's (1960) definition of unique variance contribution, suggested a procedure for weighting variables before summing to get a composite score. This method, like that of Creager and Valentine, is designed to produce desired ratios of variance contributions (but using Chase's definition of unique contribution). Guilford stated that summing several

unweighted scores causes each to contribute to the composite variance in proportion to its own variance. Then he suggested that, if we want each variable to contribute equally, we should weight each by the reciprocal of its own standard deviation. Finally, in order to achieve a given ratio of contributions to the total variance, he suggested that we multiply the initial weights by factors of importance. For example: given three variables with variances σ_1^2, σ_2^2, and σ_3^2, which we wish to contribute in the ratio $1 : 3 : 2$, we would weight them $1/\sigma_1$, $3/\sigma_2$, and $2/\sigma_3$. It can be shown however, that these weights will cause the resulting weighted variables to contribute in the ratio $1 : 9 : 4$, and that the weights we should use are $\sqrt{1} / \sigma_1$, $\sqrt{3} / \sigma_2$, and $\sqrt{2} / \sigma_3$.

The beta weights of an equation such as equation (12) may be solved for from the matrix equation

$$\beta = R^{-1}c \tag{13}$$

where: β (n by 1) is a vector of beta weights

R (n by n) is the non-singular matrix of predictor intercorrelations

c (n by 1) is a vector of correlations between each variable and the composite

Guilford gave an equation for calculating the elements of c (eqn. 16.24) which, in the notation used in this paper

would be written

$$c_i = \frac{\sum_{j=1}^{n} (r_{ij}\sigma_j)}{\sqrt{\sum_{j=1}^{n} \sigma_j^2 + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (r_{jk}\sigma_j\sigma_k)}} \qquad (14)$$

where: c_i is the correlation coefficient between variable i and the composite

r_{ij} is the correlation coefficient between variables i and j

Now, the denominator of the right hand side of (14) is equal to the standard deviation of the composite as defined by Guilford's equation 16.19. Thus equation (14) may be written

$$c_i = \frac{1}{\sigma_c} \sum_{j=1}^{n} (r_{ij}\sigma_j) \qquad (15)$$

and the matrix equation for finding all n elements of c is

$$\mathbf{c} = \frac{1}{\sigma_c} \mathbf{R} \, \boldsymbol{\sigma} \qquad (16)$$

where $\boldsymbol{\sigma}$ (n by 1) is the vector of standard deviations of the component scores. From (13) and (16)

$$\boldsymbol{\beta} = \frac{1}{\sigma_c} \mathbf{R}^{-1} \mathbf{R} \, \boldsymbol{\sigma}$$

or

$$\boldsymbol{\beta} = \frac{1}{\sigma_c} \boldsymbol{\sigma} \qquad (17)$$

Since the unique contribution of a variable has been defined as the square of the beta weight, the contribution of variable i (V_i) will be

$$V_i = \beta_i^2 = \sigma_i^2 \,/\, \sigma_c^2 \tag{18}$$

and the ratio of the contributions of two variables, i and j will be

$$\frac{V_i}{V_j} = \frac{\sigma_i^2 \,/\, \sigma_c^2}{\sigma_j^2 \,/\, \sigma_c^2} = \frac{\sigma_i^2}{\sigma_j^2} \tag{19}$$

If we use Guilford's suggested weighting method, each weight will be the product of two quantities: the reciprocal of the standard deviation of the variable, and a factor of importance for that variable. If it is desired to have variables i and j contribute in the ratio of $d_i : d_j$ then the contributions of the weighted variables will be

$$V_{i(y)} = \frac{\sigma^2_{i(y)}}{\sigma^2_{c(y)}} \qquad \text{and} \qquad V_{j(y)} = \frac{\sigma^2_{j(y)}}{\sigma^2_{c(y)}} \tag{20}$$

where the subscript (y) denotes a weighted variable, or a statistic for a weighted variable.

The variance of a weighted variable is equal to the product of the variance of the unweighted variable and the square of the weight, and thus,

$$V_{i(y)} = \frac{Y_i^2 \;\sigma_i^2}{\sigma^2_{c(y)}} \qquad \text{and} \qquad V_{j(y)} = \frac{Y_j^2 \;\sigma_j^2}{\sigma^2_{c(y)}} \tag{21}$$

and the ratio of contributions is

$$Y_i^2 \sigma_i^2 \quad : \quad Y_j^2 \sigma_j^2 \qquad (22)$$

Using Guilford's method, the two weights would be

$$Y_i = d_i / \sigma_i \qquad \text{and} \qquad Y_j = d_j / \sigma_j$$

and thus the ratio of contributions would be

$$(d_i / \sigma_i)^2 \sigma_i^2 \quad : \quad (d_j / \sigma_j)^2 \sigma_j^2$$

or

$$d_i^2 \quad : \quad d_j^2$$

Thus the variance contributions of the two variables are proportional to the squares of the factors of importance.

If, on the other hand, the weights are defined as

$$Y_i = \sqrt{d_i} / \sigma_i \qquad \text{and} \qquad Y_j = \sqrt{d_j} / \sigma_j \qquad (23)$$

the ratio of variance contributions will be

$$(\sqrt{d_i} / \sigma_i)^2 \sigma_i^2 \quad : \quad (\sqrt{d_j} / \sigma_j)^2 \sigma_j^2$$

or simply

$$d_i \quad : \quad d_j$$

Thus the variance contributions of the two variables are made to be proportional to the predetermined factors of importance.

Consider two sets of scores (two variables) having variances 16 and 9, which we wish to contribute to their composite variance in the ratio 1 : 4 . The contributions

of the unweighted variables to the composite variance (by Chase's definition) will be in the ratio 16 : 9 . Using Guilford's procedure, the weights would be 1/4 and 4/3, and the contributions would be in the ratio

$$(1/4)^2(16) \quad : \quad (4/3)^2(9)$$

or 1 : 16 . When the weights $\sqrt{1}$ / 4 and $\sqrt{4}$ / 3 , or 1/4 and 2/3 , however, are used, the contributions will be in the ratio

$$(1/4)^2(16) \quad : \quad (2/3)^2(9)$$

or 1 : 4 , which is the desired ratio.

The two definitions (Creager-Valentine and Guilford[4]) of the unique variance contribution of a part score are based on different ways of considering the composite variance. The proportion of the variance of the criterion variable associated with a set of predictor variables is given by the square of the multiple correlation coefficient (R^2). When the criterion is actually the sum (composite) of the predictors (components, or part scores), of course, perfect prediction of each criterion score is possible and the R^2 is equal to 1.00.

The coefficient of multiple correlation can be defined

[4] The above-described modification of Guilford's procedure will be referred to as the "Guilford" Procedure in the remainder of this paper

by

$$R^2 = \beta_1^2 + \beta_2^2 + \ldots + \beta_i^2 + \ldots + \beta_n^2 + 2\beta_1\beta_2 r_{12} + 2\beta_1\beta_3 r_{13} + \ldots + 2\beta_{n-1}\beta_n r_{n-1,n} \quad (24)$$

Chase (1960) defined the β_i^2 term of (24) as the unique (proportional) contribution (V_i in the notation of this paper) of the variable. The Guilford weighting method is based on this definition of the unique contribution of a variable. The weights are derived so that the unique contributions of the weighted components are in some desired ratio.

Creager and Valentine (1962), on the other hand, define the (proportional) contribution of variable i as the difference between the R^2 based upon the n component scores as predictors (1.00), and the R^2 found for the n-1 predictors left when variable i is dropped from the predictor set.

It will now be shown that the following relationships exist between the two definitions of unique contributions of variable i to the composite and the two sets of weights (Creager-Valentine = W_i ; Guilford = Y_i).

$$U_i R^{ii} = V_i \quad (25)$$

and

$$W_i^2 = R^{ii} Y_i^2 \quad (26)$$

It has already been shown (equation 18) that

$$\beta_i = \frac{\sigma_i}{\sigma_c} \tag{27}$$

and therefore

$$V_i = \beta_i^2 = \frac{\sigma_i^2}{\sigma_c^2} \tag{28}$$

and thus

$$\sigma_i^2 = V_i \sigma_c^2 \tag{29}$$

From (11)

$$\sigma_i^2 = U_i \sigma_c^2 R^{ii} \tag{30}$$

and thus from (29) and (30)

$$U_i \sigma_c^2 R^{ii} = V_i \sigma_c^2 \tag{31}$$

or

$$U_i R^{ii} = V_i$$

and thus (25) gives the relationship between the two definitions of unique contribution. As has been noted by Hazewinkel (1963), R^{ii} is the standard error of estimate for the regression of variable i on the other n-1 predictors.

It has been previously shown (pp. 46-47) that if the ratio of desired contributions of a set of n part scores is

$$d_1 : d_2 : \ldots : d_i : \ldots : d_n$$

then from (7) and (11)

$$W_i^2 = \frac{d_i R^{ii}}{\sigma_i^2} \tag{32}$$

and from (23)

$$Y_i^2 = \frac{d_i}{\sigma_i^2} \tag{33}$$

therefore from (32) and (33)

$$\frac{W_i^2}{R^{ii}} = \frac{d_i}{\sigma_i^2} = Y_i^2 \tag{34}$$

or

$$W_i^2 = R^{ii} Y_i^2$$

which proves (26).

Returning to Richardson's (1941) definition of variance contribution (pp. 43-44) it will now be shown that this definition is equivalent to what Chase (1960) called the total variance contribution of a variable. This definition arises from a different definition of the squared multiple correlation coefficient than that of (24), that is

$$R^2 = \beta_1 c_1 + \beta_2 c_2 + . . . + \beta_i c_i + . . . + \beta_n c_n \tag{35}$$

Since $R^2 = 1.0$ when the "criterion" is a linear composite of the unweighted predictors, the $\beta_i c_i$ terms in (35) sum

to unity and thus Chase defines such terms as the total variance contributions of the respective variables to the composite. Richardson's definition of the variance contribution of variable i (p. 44) was

$$\sigma_i \sum_{j=1}^{n} (r_{ij}\sigma_j) \qquad (36)$$

Where the n such terms sum to σ_c^2 . Thus the proportional contribution of variable i will be

$$\frac{\sigma_i}{\sigma_c^2} \sum_{j=1}^{n} (r_{ij}\sigma_j) \qquad (37)$$

Consider Chase's definition of total contribution of variable i as T_i in the expression

$$T_i = \beta_i c_i \qquad (38)$$

From (15) and (38)

$$T_i = \frac{\beta_i}{\sigma_c} \sum_{j=1}^{n} (r_{ij}\sigma_j) \qquad (39)$$

and from (17) or (18)

$$T_i = \frac{\sigma_i}{\sigma_c^2} \sum_{j=1}^{n} (r_{ij}\sigma_j) \qquad (40)$$

which is equal to (37). Thus Richardson's definition of variance contribution is equivalent to Chase's definition of the total contribution of a variable. It has been pointed out both by Chase and by Ward (1962) that this

quantity

> has the disadvantage of not yielding a break-down between variance which is directly associated with the ith variable, and that portion of the variance shared among the several independent variables which is to be associated with the ith variable (Chase, 1960, p. 265).

Horst (1961a; 1961b; 1965, pp. 569-576) has developed a procedure for weighting several sets of variables so as to yield the highest possible sum of correlation coefficients between the weighted composite scores of each set. The method is similar to canonical correlation, but extended to more than two sets of variates.

Consider three sets of variables, and the composites found by summing each set, for example

$$\begin{aligned} X_c &= X_1 + X_2 + \ldots + X_{nx} \\ Y_c &= Y_1 + Y_2 + \ldots + Y_{ny} \\ Z_c &= Z_1 + Z_2 + \ldots + Z_{nz} \end{aligned} \tag{41}$$

where nx, ny, and nz are the numbers of variables in the respective sets. The problem is to find three sets of weights which, when applied to the three sets of variables, will yield the highest possible coefficients of correlation between the weighted linear composites of the form

$$\begin{aligned} X_{c(d)} &= d_1X_1 + d_2X_2 + \ldots + d_{nx}X_{nx} \\ Y_{c(e)} &= e_1Y_1 + e_2Y_2 + \ldots + e_{ny}Y_{ny} \\ Z_{c(f)} &= f_1Z_1 + f_2Z_2 + \ldots + f_{nz}Z_{nz} \end{aligned} \tag{42}$$

The following is a brief resume of the Horst procedure as used in this study. More detailed descriptions and proofs may be found in the writings of Horst (1961a, 1961b, 1965). It should be pointed out that Horst has presented four procedures for weighting m different sets of data, each designed to produce a specific relationship between the linear composites. The procedure employed in this study is that named the "maximum correlation method" (Horst, 1965, p. 566), which is most completely described in Horst's first (1961b) article on these "multiple set factor analysis" procedures. This particular technique was chosen because the weights are derived so as to maximize the sum of the correlations between the weighted composites.

Three basic score matrices are defined for each set of data, as well as three supermatrices (43, 44, and 48 below) formed by placing the basic score matrices of each set beside each other. Following Horst's notation, the supermatrices will be denoted by capital letters without subscripts (X, Y, Z, etc.) while the matrices of a given set will be denoted by capital letters preceeded by a subscript denoting the set (${}_1X$, ${}_1Y$, ${}_1Z$, etc.). The three basic score matrices are

$$X = ({}_1X, {}_2X, \ldots {}_mX) \tag{43}$$

= the supermatrix of m sets of unweighted scores in standard form

$$U = ({}_1U, {}_2U, \ldots {}_mU) \tag{44}$$

= a supermatrix formed by transformations of the ${}_iX$ matrices of the form

$${}_iU = {}_iX \, t_i^{\prime -1} \tag{45}$$

where t_i is a triangular matrix formed by a square root factoring of the matrix of intercorrelations within set i. That is

$$t_i t_i^{\prime} = G_{ii} \tag{46}$$

where

$$G_{ii} = \frac{{}_iX^{\prime} \, {}_iX}{N} \tag{47}$$

and

$$Z = ({}_1Z, {}_2Z, \ldots {}_mZ) \tag{48}$$

= a supermatrix of transformations of the ${}_iX$ matrices of the form

$${}_iZ = {}_iU \, {}_i\beta = {}_iX \, {}_ib \tag{49}$$

where ${}_i\beta$ and ${}_ib$ are matrices of weights derived so as to maximize the sums of the correlations between weighted linear composites.

From (45) and (49) it can be seen that

$${}_i\beta = t_i^{\prime} \, {}_ib \tag{50}$$

Because of the criterion for deriving the weights, the

supermatrix of correlations between the scores of the ${}_1Z$ matrices, ρ , where

$$\rho = \frac{Z' Z}{N} \tag{51}$$

consists of diagonal submatrices that are identity matrices and off-diagonal submatrices that contain the correlation coefficients to be maximized. As in canonical correlation, successive sets of weights may be found that will give several orthogonal solutions to the problem. The diagonal submatrices of ρ contain the correlation coefficients between pairs of orthogonal composites and are thus identity matrices. In the present study only one set of weights was derived, that which would yield the highest possible sum of correlations between composites.

Consider the supermatrix below (52), resulting from the application of the method to three sets of variables ($m = 3$) each set containing at least three variables (the numbers of variables need not be identical in each set), and three orthogonal sets of linear composites having been derived (as in canonical correlation, the maximum number of new variates which may be derived is equal to the rank of the intercorrelation matrix of lowest rank among the m such original matrices).

$$\rho = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ab_{11} & ab_{12} & ab_{13} & ac_{11} & ac_{12} & ac_{13} \\ a_{21} & a_{22} & a_{23} & ab_{21} & ab_{22} & ab_{23} & ac_{21} & ac_{22} & ac_{23} \\ a_{31} & a_{32} & a_{33} & ab_{31} & ab_{32} & ab_{33} & ac_{31} & ac_{32} & ac_{33} \\ ab_{11} & ab_{12} & ab_{13} & b_{11} & b_{12} & b_{13} & bc_{11} & bc_{12} & bc_{13} \\ ab_{21} & ab_{22} & ab_{23} & b_{21} & b_{22} & b_{23} & bc_{21} & bc_{22} & bc_{23} \\ ab_{31} & ab_{32} & ab_{33} & b_{31} & b_{32} & b_{33} & bc_{31} & bc_{32} & bc_{33} \\ ac_{11} & ac_{12} & ac_{13} & bc_{11} & bc_{12} & bc_{13} & c_{11} & c_{12} & c_{13} \\ ac_{21} & ac_{22} & ac_{23} & bc_{21} & bc_{22} & bc_{23} & c_{21} & c_{22} & c_{23} \\ ac_{31} & ac_{32} & ac_{33} & bc_{31} & bc_{32} & bc_{33} & c_{31} & c_{32} & c_{33} \end{bmatrix}. \qquad (52)$$

Supermatrix ρ consists of six submatrices that are distinct (the matrices below the major diagonal are identical to those above the diagonal). Thus ρ can be considered as

$$\rho = \begin{bmatrix} A & AB & AC \\ AB & B & BC \\ AC & BC & C \end{bmatrix} \qquad (53)$$

Columns 1, 4, and 7 of (52) contain the correlations involving the first set of maximally correlated variates, columns 2, 5, and 8 contain correlations involving the second set of variates, and columns 3, 6, and 9 contain correlations involving the third set. Since successive linear composites within sets are orthogonal, A, B, and C in (53) are identity matrices. In this example the first set of weights is derived so as to maximize the sum defined in

(54) where the terms on the right are defined in (52).

$$\phi_1 = ab_{11} + ac_{11} + bc_{11} \tag{54}$$

The second set of weights maximizes

$$\phi_2 = ab_{22} + ac_{22} + bc_{22} \tag{55}$$

and the third maximizes

$$\phi_3 = ab_{33} + ac_{33} + bc_{33} \tag{56}$$

The off-diagonal elements of AB, AC, and BC in (53) are not necessarily zero, that is, the first derived variate of the first set of variables is not necessarily orthogonal to the second and third derived variates for the second and third sets of variables.

In the present study, as previously mentioned, only one set of new variates was derived, and thus each submatrix of (53) contains only one element, that is

$$\rho = \begin{bmatrix} 1.0 & ab_{11} & ac_{11} \\ ab_{11} & 1.0 & bc_{11} \\ ac_{11} & bc_{11} & 1.0 \end{bmatrix} \tag{57}$$

where ab_{11}, for example, is the correlation coefficient between the weighted linear composites of the first two sets of variables. It may be noted that being "generalized canonical correlations" (Horst 1961a) the coefficients of (57) are all positive numbers. Horst (1961b) presents an iterative procedure, best described as a generalization of

Hotelling's (1935, 1936) iterative solution to the canonical correlation problem. Although Horst has not developed a rigorous proof that the solution will converge on stable sets of weights with functions like ϕ_1, ϕ_2, and ϕ_3 of (54), (55), and (56) maximized, he showed (1961b, pp. 145-146) that the solution is equivalent to Hotelling's solution for two sets of variables and pointed out that it can be proved that the latter converges to the largest latent root. He also mentioned that "experimental results so far obtained have converged (1961b, p. 147)."

It may be seen from the literature cited above that there is a lack of agreement among the different writers as to what is meant by the contribution of a predictor variable to a criterion variable, or a part score to a composite. Another indication of the disagreement may be seen in a recent controversy between Hoffman (1960, 1962) and Ward (1962, 1963).

Hoffman (1960) discussed a "relative weight" which is equivalent to Chase's (1960) definition of the total variance contribution of a variable. A major disadvantage of this definition has been previously discussed (p. 58). Hoffman's definition, in the notation of this paper would be defined for variable i as

$$\frac{\beta_i c_i}{R_c^2} \qquad (58)$$

and is thus equivalent to Chase's definition of total contribution divided by the coefficient of multiple correlation between the criterion and all predictors. Since R_c^2 would be a constant for all predictors, Hoffman's relative weights are in the same ratio as are Chase's total contributions. When the "criterion" is the sum of the "predictors" as are the variables used in this study, of course, $R_c^2 = 1.0$, and the two definitions are equivalent.

Ward, on the other hand, appears to favor the use of the concept of unique contribution as defined by Creager and Valentine (1962), discussed above.

Gibson (1962), discussed the problem and suggested that the difference of opinion could be eliminated by deriving orthogonal variables from the predictor variables, and thus one could work with a set of variates for which the two definitions of contribution would be numerically equivalent. Both Gibson and Kaiser (1967) have derived methods of orthogonalizing variable sets such that the orthogonal variables are maximally correlated with the original variables. However, as is pointed out by Ward (1963), researchers may not wish to deal with artificially orthogonalized variables when they could use the original ones, which are likely to be more meaningful to them. It can also be argued that if characteristics of the population

being studied are correlated they should be investigated as they are, that is, as correlated variables.

Of the two definitions of the unique contribution of a variable to a composite (that of Chase, and that of Creager & Valentine), there appears to be no mathematical basis for choosing either as more appropriate than the other. As a matter of interest, however, it might be pointed out that the Creager-Valentine definition is that usually used when referring to the contribution of a predictor in regression analysis. When the least squares procedure is used to estimate the parameters in the fixed-effects-model analysis of variance a similar procedure is used to partition the variance attributable to the different effects. Thus the Creager-Valentine definition of the variance contribution of a predictor to the criterion is probably more widely used than is the Chase definition.

CHAPTER III

PROCEDURES

The writer has been unable to find any applications of the above mentioned weighting systems in studies of written composition. It was proposed that the Creager-Valentine (1962) and Guilford (1965) methods could be used to weight the part scores derived from the analytic marking of essays, in order that the weighted scores may be actually contributing to the composite variance in predetermined ratios. The weighting methods may be used in such a way that not only are the overall variance contributions of the part scores brought into a desired ratio, but if the method is used separately on the ratings of each reader and differential weights (different weights for each reader) are used, the grades of the various readers on the same paper should be more similar. It was thought possible that such differential weighting of each marker's scores could result in increased inter-marker reliability coefficients. In addition it was decided to use Horst's (1961a, 1961b, 1965) procedure in order to investigate the nature of weighting systems that would maximize the inter-marker reliability coefficients.

<u>The Sample</u>

In 1964, Nyberg (1966) randomly selected 103 essay

papers, written on the same topic, from the papers of 12,921 Alberta grade 12 students writing the provincial departmental examination in English 30. Students had a choice of two topics. The essays were marked by an analytical procedure with 27 markers grading each of the 103 papers for mechanical errors, and 21 different markers grading each for style and content.

In the style-content grading procedure, marks were given for a total of 16 criteria as shown in Table 2.

In the mechanics grading procedure, each paper was checked for six types of errors: spelling, punctuation, word usage, grammar, sentence errors, and form. On each of these variables either two or three marks were deducted from the total essay score, with a predetermined limit on the number of marks that could be deducted for each of the variables, and a maximum total of 125 marks that could be deducted from a given paper's score. The marking variables are shown in Table 3.

Three criteria are used as a basis for selection of readers by the Alberta Department of Education. The readers actually used consist of paid volunteers who meet these criteria. The criteria, as they apply to the English 30 paper, are that the individual holds a degree from a recognized university, has a permanent teaching certificate that is valid in the province of Alberta (which implies

Table 2

Style-content Grading Variables

Variable	No. Marks
Material	
1. Significance	10
2. Relevance	10
3. Originality	5
Organization	
4. Plan	5
5. Relation of plan to essay	5
6. Introduction	5
7. Order	5
8. Emphasis	5
9. Conclusion	5
Diction	
10. Exactness and vividness of nouns, verbs, adjectives, adverbs	10
11. Figures of speech, comparisons, illustrations	5
12. Breadth of vocabulary	5
Sentence Organization	
Variety	
13. structure	15
14. beginnings	5
15. Economy	15
16. Total Impression	15
Total	125

Table 3

Mechanics Grading Variables

Variable	Marks per Error	Maximum
1. Spelling	3	39
2. Punctuation	2	26
3. Word Usage	2	24
4. Grammar	3	33
5. Sentence Errors	3	39
6. Form	2	14
	Total	175 marks
	Maximum allowable deduction	125 marks

that he has had at least two years of successful teaching experience as judged by a high school inspector), and has taught English 30 during the school year just completed at the time of the marking session.

The data from Nyberg's study was used in the present study and similar data was collected from the 1967 Alberta grade 12 departmental examinations in English 30. Five mechanics markers and two style-content markers who took part in Nyberg's study were found to be again grading essays in 1967. At the time the 1967 data were collected, the papers were all marked and filed according to candidates numbers. Sections of the files were randomly selected and all essays in these sections that were read by the seven markers were examined and part scores recorded. As a result of

this sampling technique, the number of papers included in the 1967 data varied from reader to reader.

Although the marking criteria employed in 1964 and 1967 were essentially the same, it was found that the style-content markers had been instructed to pool some of the criteria in 1967. As a result there were only 10 style-content part scores as compared to 16 in the 1964 data. In order to make comparisons between data from the two years, it was therefore necessary to similarly pool the part scores from the 1964 data. No such problem arose with the mechanics part scores.

In the 1967 marking session, only one reader graded each paper, so it was impossible to calculate inter-marker reliability coefficients. The purpose of collecting this additional data was to examine the stability of the Creager-Valentine and Guilford weights over time for each marker. It was also possible to examine the stability of weights over topics graded by the same marker in the same year because, in 1967, candidates were instructed to choose one of four topics for their essays.

Analysis of the Data

Procedures used in this study that are not standard statistical techniques are described in Appendix A. The computing algorithms used, along with their derivations or

sources are presented in the appendix, and are numbered so as to match the numbers in the outline of procedures below.

In deriving the Creager-Valentine and Guilford weights, the ratio of desired variance contributions of part scores was defined as the ratio of the total possible scores on each marking criterion as reported in Tables 2 and 3.

1. Reliability of Mechanics Markers

1.1 Creager-Valentine weighting system

1.1.1 For each marker, the unique contribution (Creager-Valentine definition) of each part score to the composite variance was calculated

1.1.2 The appropriate Creager-Valentine weights were calculated for each part score of each marker

1.1.3 The raw part scores of each essay for each marker were summed and intercorrelations between markers' total mechanics scores calculated. The average of these reliability coefficients between the markers was computed by two methods: finding the mean of the coefficients; and using Fisher's z transformation (Guilford, 1965, pp. 348-349) on all coefficients, finding the mean z, and transforming the mean z back to a correlation coefficient.

1.1.4 The part scores of each paper for each marker were weighted by the Creager-Valentine weights and average inter-marker reliability coefficients calculated

as in 1.1.3.

1.2 Guilford weighting system

1.2.1 For each marker, the unique contribution (Guilford-Chase definition) of each part score to the composite variance was calculated

1.2.2 The appropriate Guilford weights were calculated for each part score of each marker

1.2.3 The part scores of each essay for each marker were weighted by the Guilford weights and average inter-marker reliability coefficients of the 21 markers calculated as in 1.1.3

1.3 Horst weighting system

1.3.1 Appropriate Horst weights were calculated for a random sample of 16 of the 27 mechanics markers. Since the Horst procedure extracts standard score (z score) weights, the weights were converted to raw score weights by dividing each by the standard deviation of the appropriate part score.

1.3.2 Average reliability coefficients for the weighted scores were calculated as in 1.1.3

2. Reliability of Style-content Markers

The procedures used were identical to those used for the mechanics markers with one exception. Two random samples of six markers each were drawn for the Horst weighting

system and independently analyzed.

3. Double Weighting Procedures

3.1 Each of the 16 mechanics markers scores were weighted by the Horst weights, then new Creager-Valentine and Guilford weights were calculated from the weighted scores. The average coefficients for both sets of double-weighted scores were calculated as in 1.1.3.

3.2 The procedures in 3.1 were repeated independently for both samples of 6 style-content markers

4. Stability of Creager-Valentine and Guilford Weights

Creager-Valentine and Guilford weights were calculated for each marker's scores for each topic. Since little is known about the distribution of such sets of weights, each set was ranked and rank order correlational procedures were used to estimate the stability. The following procedures were used for both the Creager-Valentine and the Guilford weights.

4.1 Mechanics markers: stability of weights within markers over topics marked at the same time

4.1.1 A coefficient of concordance (W) (Ferguson, 1966, pp. 225-228) was calculated within each of the five markers between the ranks of the weights on the four 1967[5]

[5] A number used to denote an essay topic refers to the year in which the topic was assigned

essay topics.

4.1.2 A W was calculated between the markers for each of the four 1967 topics.

4.1.3 It was reasoned that an indication of the stability of the weights over topics marked at the same time could thus be found by comparing the median within-marker W to the median between-marker W. If a set of weights derived for a marker are stable there should be a higher median W within markers than between markers.

4.2 Mechanics markers: stability of weights within markers over topics marked at two different times

4.2.1 A W was calculated within each of the five markers between the ranks of the weights on all five topics (4 from 1967, 1 from 1964).

4.2.2 A W was calculated between the markers for each of the five topics

4.2.3 The median within marker W was compared to the median between marker W

4.3 Style-content markers: stability of weights within markers over topics marked at the same time. The procedures in 4.1 were repeated with the style-content markers who marked both in 1964 and in 1967. Since there were only two such markers the results were not used to make any inferences.

4.4 Style-content markers: stability of weights within

markers over topics marked at two different times. The procedures in 4.2 were repeated with the two style-content markers who marked both in 1964 and 1967.

5. Effects on gross grade categories of adjustments to equate all markers' means.

5.1 Mechanics markers

5.1.1 The mean (grand mean) and the standard deviation of the distribution of all scores of all markers (103 x 27 scores) were calculated

5.1.2 The mean of each marker's scores on the 103 papers was calculated

5.1.3 A distribution of adjusted scores (103 x 27) was found by adding to the scores of each marker the difference between the grand mean and that marker's mean

5.1.4 The standard deviation of the 103 x 27 adjusted scores was calculated and the scores of the two distributions (raw and adjusted) were transformed to z scores

5.1.5 All scores were classified in gross score categories as follows

$$2.0 \leq z$$

$$1.0 \leq z < 2.0$$

$$0.0 \leq z < 1.0$$

$$-1.0 \leq z < 0.0$$

$$-2.0 \leq z < -1.0$$

$$z < -2.0$$

5.1.6 For each marker, the number of individual essay scores that shifted from one gross classification to another when the scores were adjusted was calculated and converted to a proportion

5.1.7 Using the method described by Ferguson (1966, p. 158), 95% confidence limits were established for the proportions in order to determine whether or not the proportions differred from zero at the .05 significance level

5.2 Style-content markers

The procedures in 5.1 were repeated with the style-content markers

6. Effects on gross grade categories of the three weighting systems

6.1 Mechanics markers

6.1.1 Creager-Valentine weights were applied to the part scores of each marker and a total weighted score calculated

6.1.2 The resulting scores were adjusted in order to equate markers' means

6.1.3 The scores were then classified into gross score categories as in 5.1.5 and proportions of shifts in categories upon weighting were calculated

6.1.4 The significance of the proportions was tested as in 5.1.7

6.1.5 The procedures in 6.1.1 through 6.1.4 were repeated with the Guilford weights, and then with the Horst weights

6.2 Style-content markers

The procedures in 6.1 were repeated with the marks of the style-content markers

CHAPTER IV

RESULTS

Effects of Weighting Systems on Inter-marker Reliability

For both the 27 mechanics markers and the 21 style-content markers, it was found that weighting each marker's part scores by the Creager-Valentine weights, or by the Guilford weights, before summing the part scores, resulted in sets of composite scores between which inter-marker correlation coefficients averaged slightly lower than did the similar coefficients for the unweighted scores. In other words, the average inter-marker reliability was slightly decreased by applying the Creager-Valentine or the Guilford weights.

It was also found that sets of weights could be found for each marker by the use of the Horst procedure, that when applied to each marker's part scores, resulted in total scores that, on the average, were more highly correlated than were the unweighted scores. That is, for random samples of sixteen mechanics markers, or six style-content markers, the average reliability coefficient was increased by the application of Horst weights.

On the other hand, it was found that the use of two of the weighting procedures (Creager-Valentine & Horst, or Guilford & Horst), in an attempt both to achieve a desired

ratio of variance contributions of part scores to the total score and to maximize the average inter-marker reliability coefficients, resulted in average reliability coefficients that were lower than average raw-score reliability coefficients for each of the three samples of markers.

It was found, in all cases, that the use of the z transformation of the correlation coefficients prior to averaging, and then converting the average z back to an r, resulted in a higher coefficient than was found by simply taking the mean of the r's. The difference, however, was never greater than .02 and it was decided to report herein only the more conservative estimates found by taking the mean inter-marker reliability coefficients without using the z transformation. The coefficients are reported in Table 4.

The values of the weights found from the data are reported in Appendix B. It may be noted that, although it was found possible to increase the average inter-marker reliability coefficient for mechanics markers from .76 to .82 through the use of the Horst procedure, this coefficient results from a large weight being applied to each marker's spelling score, and relatively small weights being applied to each of his other part scores. The validity of the resulting scores is therefore rather

Table 4

Average Inter-marker Reliability Coefficients

Scores	Markers				
	All 27 Mech.	Sample of 16 Mech.	All 21 Style	Sample of 6 Style	Sample of 6 Style
Raw	.769	.763	.595	.569	.553
Creager-Valentine weighted	.719	.713	.588	.555	.542
Guilford weighted	.721	.721	.592	.562	.547
Horst weighted	--	.816	--	.736	.690
Horst & Creag. weighted	--	.578	--	.487	.369
Horst & Guilf. weighted	--	.580	--	.491	.358

questionable. As would be anticipated under such circumstances, the spelling scores contributed much more to the composite score variance than did the other part scores.

The unique variance contribution of a part score to the total score by either the definition of Creager and Valentine, or that of Chase, is proportional to the variance of the part score (see equation 11, p. 47, and equation 18, p. 51). Therefore an examination of the Horst weights to be applied to the standard (z) part scores, which all have a variance of one, provides a better indication of the difference between the contributions of

spelling and the other part scores, than does examination of the raw score weights. The z-score weights are reported in Table B4 of Appendix B, while the raw score weights are reported in Table B3 of the appendix.

Because of the previously mentioned very high weighting given spelling scores by the Horst procedure, it was decided to repeat the Horst procedure for the sample of 16 mechanics markers, with the spelling scores omitted, thus reducing the number of part scores to 5. When this procedure was carried out, it was found that the Horst method could produce weights that would increase the average inter-marker reliability coefficient from .627 to .647. The weights so derived showed no systematic similarity across all markers as was shown when the spelling scores were included. Neither was any such systematic similarity between part score weights found in the style-content markers' Horst weights. Rather, it was found that for different markers, different part scores would be most heavily weighted. The z score weights are reported in Appendix B.

As might be expected, the particular part score weighted most heavily for a given marker was that in which he was most in agreement (as determined by inter-marker correlation coefficients of part scores) with his colleagues. In addition, the part scores that were most heavily weighted for a given marker also contributed more to the weighted

composite variance than did the other part scores.

Stability of Creager-Valentine and Guilford Weights

As was expected, the median coefficient of concordance (w) within markers between topics marked at the same time (1967) was found to be higher than was the median W between markers within topics, for the five mechanics markers who marked both in 1964 and in 1967, and for both the Creager-Valentine and the Guilford weights. Thus an indication that these weighting systems show some stability within markers over topics was shown. It must be admitted, however, that the method of estimating stability is rather weak.

On the whole, there appeared to have been no tendency for either the Creager-Valentine or the Guilford weights to show greater stability, as defined herein, than did the other. The W's for the Creager-Valentine weights were higher than those for the Guilford weights of some markers and some topics, but some were also lower.

A summary of data used to estimate the stability of mechanics markers is reported in Table 5. Although similar calculations were made on the style-content markers' data, there were only two markers reading essays for style and content in both 1964 and 1967, and it was felt that data based upon such a small sample of markers would be

unreliable. The resulting data are therefore reported in Appendix C and will not be discussed here.

Table 5

Coefficients of Concordance--Topics Marked at Same Time

Marker No.	Within Marker W's		Topic No.	Between Marker W's	
	C-V wts.	G wts.		C-V wts.	G wts.
6	.779	.821	1	.502	.515
8	.907	.857	2	.561	.575
12	.764	.714	3	.721	.657
26	.757	.657	4	.383	.643
27	.564	.693			
Median	.76	.71		.53	.61

When similar calculations were made in an attempt to look at the stability of the Creager-Valentine and the Guilford weights over both time and topics, the results were very similar. The data are reported in Table 6.

Perhaps a better estimate of the stability of weights over time might be found by calculating a Spearman rank order correlation coefficient between each of the 1967 topics and the 1964 topic for each marker, and then finding the median of the 20 resulting coefficients. When such calculations were made, the median coefficient was found to be .60 for both the Creager-Valentine and the Guilford

weights. In order that these coefficients could be compared to the W's reported in Tables 5 and 6, Siegel's (1956, p. 229) equation 9.14 expressing the relationship between an average Spearman rho and a coefficient of concordance was used. A rho of .60 was found to be equal to a W of .80.

Table 6

Coefficients of Concordance--All Topics

Marker No.	Within Marker W's		Topic No.	Between Marker W's	
	C-V wts.	G wts.		C-V wts.	G wts.
6	.726	.717	67-1	.502	.515
8	.826	.790	67-2	.561	.575
12	.771	.707	67-3	.721	.657
26	.680	.621	67-4	.383	.643
27	.598	.721	64	.451	.653
Median	.73	.72		.50	.64

Effects on Gross Grade Categories of Adjustments to Equate all Markers' Means

Using the method described by Ferguson (1966, p. 158) it was determined that for an N of 103 any proportion greater than .05 differs significantly from zero at the .05 level. Using this criterion, there would be only two markers out of the 27 mechanics markers and two out of the 21 style-content markers for whom there were not significant

proportions of individual essay scores shifting from one gross score category to another when scores were adjusted in order to equate markers' means. When a more conservative proportion (.10) was used as a criterion for proportion of shifts that was considered to be of practical significance, there were still only 5 mechanics and 8 style-content markers with non-significant proportions of shifts. The data are reported in Table 7.

Table 7

Frequencies of Score Shifts when Adjusted--by Proportions

Proportion of Shifts	Frequency of Mechanics Markers	Frequency of Style Markers
.70 - .79	1	
.60 - .69		
.50 - .59	3	
.40 - .49	3	1
.30 - .39	1	
.20 - .29	2	2
.10 - .19	12	10
.00 - .09	5	8
N	27	21

Effects of Weights on Gross Grade Categories

For all of the mechanics markers, the proportion of shifts in gross grade categories was at least .10 when scores were weighted by either the Creager-Valentine or

the Guilford methods. For the Style-content markers, however, there were three markers having proportions of less than .10 when the Creager-Valentine weights were applied and six markers having such proportions when the Guilford weights were applied. The data are reported in Table 8.

Table 8

Frequencies of Score Shifts under weighting--by Proportions

Proportion Of Shifts	Mechanics Markers		Style-content Markers	
	C-V wts.	G wts.	C-V wts.	G wts.
.70 - .79	1	1		
.60 - .69	1			
.50 - .59	4	3		
.40 - .49	3	3	1	1
.30 - .39	3	1	1	
.20 - .29	8	8	3	2
.10 - .19	7	11	13	12
.00 - .09			3	6
N	27	27	21	21

When the Horst weights were applied to the random sample of mechanics markers, none of the sixteen markers had proportions of shifts lower than .10 and only three had a proportion of less than .30. For the two style-content samples, there were no markers who had proportions of shifts lower than .30. These data are reported in Table 9.

Table 9

Frequencies of Score Shifts
for Horst Weights--by Proportions

Proportion Of Shifts	Mechanics Markers	Style-content Markers	
		Sample 1	Sample 2
.70 - .79		1	
.60 - .69		2	
.50 - .59	2	2	
.40 - .49	2	1	2
.30 - .39	9		4
.20 - .29	2		
.10 - .19	1		
N	16	6	6

One of the interesting findings with respect to the effects of adjusting the scores in order to equate markers means, and weighting part scores in order to obtain desired ratios of variance contributions, was the close agreement between the different transformations. That is, the proportions of shifts were very similar under all three transformations used with the complete samples of markers. The proportions of shifts for each marker, along with proportions in which similar shifts resulted under two different transformations are reported in Appendix D.

CHAPTER V

DISCUSSION AND IMPLICATIONS

Differential Weighting of Part Scores Assigned by Analytic Essay Readers

The results of this study suggest that, for the kinds of readers and compositions involved, differential weighting designed to produce desired ratios of contributions of part scores to total score variance, slightly reduces the average inter-marker reliability coefficient. It is the opinion of the writer, however, that the validity of the resulting scores is improved, because the total scores after weighting represent predetermined desired combinations of part scores. That is, it had been decided that there were certain criteria by which the standards of the individual compositions were to be judged, and that a certain amount of the total possible score should be derived from each marking criterion. If then, the part scores derived from these criteria are weighted so that, for each marker, the contribution of each part score is a desired proportion of the total variance, the resulting total scores should be a better representation of the desired measurement of composition skills.

Because of the widespread belief that the reliability coefficient of a set of measurements <u>always</u> sets the upper

bound on the validity of the measurements, the discussion in the above paragraph may be a source of concern to some readers. The writer, however, tends to agree with Torgerson and Green (1952) who stated that

> The usual operational definition of validity as "the correlation of a test with a criterion," while reasonably adequate for determining the validity of an aptitude test, is inapplicable when considering measures of achievement. The achievement test is actually a measure of the criterion itself. This implies that the statistical techniques which have proved to be valuable in increasing the correlational validity of aptitude tests should not be applied uncritically to the achievement test field. The selection or rejection of items on the basis of the usual item analysis data is a case in point. While this method will very likely increase the correlational validity of a test, the effect on the validity of the test as a measure of achievement is doubtful (p. 355).

On the other hand, it is the opinion of the writer, that scores resulting from the use of weights derived by the Horst procedure result in decreased validity. The procedure results in weights that produce a higher average inter-marker reliability coefficient, but in order to achieve such a coefficient, part scores are weighted in accordance with their own inter-reader reliabilities. Thus, if the only criterion upon which essay readers are in agreement is spelling, the variance of weighted total scores will be based mainly on spelling.

Methods such as the Horst procedure used in this study may be valuable for the studying of the kinds of marking criteria upon which readers are able to show the most

agreement, rather than for differential weighting of scores.

It would appear from the data of this study that the achievement, by differential weighting, of both high intermarker reliability coefficients, and desired proportions of variance contributions by part scores to total scores, are incompatible aims with respect to essay scores. However, since this is the first study of this nature, some of the methods were rather crude, and it is possible that refinements of some of the procedures could result in other conclusions. It should also be pointed out that the results found in this study may be specific to data matrices having the same characteristics as those of this study.

Differences between the Creager-Valentine and Guilford procedures used in this study appear to be slight, and neither definition of variance contribution and accompanying derived weights could be concluded to be more acceptable for purposes such as those of this study than is the other.

Before any more general conclusions could be made with respect to the procedures used in this study it would be desirable to study these procedures with different sets of readers and compositions.

Stability of Creager-Valentine and Guilford Weights

The part of this study in which the techniques used

were probably the most crude was in the estimation of the stability of the weights. Further studies of the nature of the distributions of such weights would be invaluable to study of their stability, in that more sophisticated techniques could then perhaps be used for estimating the stability.

The only conclusion possible with the present data is that some indication of stability, as defined herein, was found with the kind of readers and compositions used.

Effects of Transformations on Individual Scores

This study showed that the use of differential weighting procedures, and even the use of simple transformations to equate markers means, can affect a significant proportion of individuals' essay grades. In the light of this finding it might be appropriate for those people who are working in the field of measurement of composition skills, or skills measurable only through written composition, to consider the validity of the transformed scores as compared to that of the raw scores.

Perhaps one of the more interesting findings of this study is the apparent similarity between the shifts in grades that took place under different kinds of transformations. For some purposes, the more simple transformation, equating markers' mean assigned scores, may be sufficient

for obtaining the kind of measurement desired. Before any generalizations are possible about effects on individual scores, it would be desirable to investigate the shifts in scores under different conditions such as: different essay topics, different kinds of subjects, different kinds of readers, different marking criteria, and perhaps most important different sizes of correlation coefficients between the readers, both in overall grades, and in part scores.

The main purpose of this study was to study some new techniques for examining some old problems related to measurement by the essay examination. It is the writer's opinion that, although a study such as this may involve the use of some rather crude techniques, it is a necessary first step in the evaluation of new techniques. Before any suggestions can legitimately be made with respect to techniques such as those used in this study, more detailed study of certain aspects are necessary. However, in order to determine whether or not an extensive examination of some of the details of a procedure are warranted, some investigation of the value of the technique as a whole should be undertaker. Within this context, it is the writer's opinion that this study has shown that the Creager-Valentime and Guilford weighting procedures have sufficient merit to warrant further, more extensive investigation. The Horst procedure may also prove valuable, especially

in the study of marker characteristics, but the writer would question its use for differential weighting of essay scores because of the previously mentioned limitations with respect to validity.

REFERENCES

REFERENCES

Anderson, C. C. The new STEP Essay Test as a measure of composition ability. Educational and Psychological Measurement, 1960, 20, 95-102.

Ballard, P. B. The new examiner. London: Hodder & Stoughton, 1925. (cited in Cast, 1939, p. 258).

Bracht, G. H., & Hopkins, K. D. Objective & essay tests: do they measure different abilities? Paper presented at joint session of annual meetings of American Educational Research Association and National Council for Measurement in Education, Chicago, February, 1968.

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in written composition. Champaign, Illinois: National Council of Teachers of English, 1963.

Britton, J. Experimental marking of English compositions written by fifteen-year-olds. Educational Review (Birmingham), 1963, 16, 17-23.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. (Ed.). Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 171-246.

Cast, B. M. D. The efficiency of different methods of marking English composition, part I. British Journal of Educational Psychology, 1939, 9, 257-269.

Cast, B. M. D. The efficiency of different methods of marking English composition, part II. British Journal of Educational Psychology, 1940, 10, 49-60.

Chase, C. I. Computation of variance accounted for in multiple correlation. Journal of Experimental Education, 1960, 28, 265-266.

Coffman, W. E., & Kurfman, D. G. Single score versus multiple score reading of the American History Advanced Placement Examination. College Entrance Examination Board Research and Development Report RDR-65-6, No. 14. Princeton: Educational Testing Service, 1966.

Coward, A. F. A comparison of two methods of grading English compositions. Journal of Educational Research, 1952, 46, 81-93.

Creager, J. A., & Valentine, L. D., Jr. Regression analysis of linear composite variance. Psychometrika, 1962, 27, 31-37.

Diederich, P. B. How to measure growth in writing ability. English Journal, 1966, 55, 435-449.

Edgerton, H. A., & Kolbe, L. E. The method of minimum variation for the combination of criteria. Psychometrika, 1936, 1, 183-187.

Ferguson, G. A. Statistical analysis in psychology and education. New York: McGraw-Hill, 1966 (second edition).

Finlayson, D. S. The reliability of the marking of essays. British Journal of Educational Psychology, 1951, 21, 126-134.

Gibson, W. A. Orthogonal predictors: a possible resolution of the Hoffman-Ward controversy. Psychological Reports, 1962, 11, 32-34.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1965 (fourth edition).

Guttman, L. Mathematical and tabulation techniques. Supplementary study B in Horst, P. (Ed.). The prediction of personal adjustment. Social Science Research Council Bulletin 48. New York: SSRC, 1941, pp. 251-364.

Hazewinkel, A. A note concerning the Creager-Valentine paper. Psychometrika, 1963, 28, 105-108.

Hoffman, P. J. The paramorphic representation of clinical judgment. Psychological Bulletin, 1960, 57, 116-131.

Hoffman, P. J. Assessment of the independent contributions of predictors. Psychological Bulletin, 1962, 59, 77-80.

Holzinger, K. J., & Harman, H. H. Factor analysis, a synthesis of factorial methods. Chicago: University of Chicago Press, 1941.

Horrocks, J. E. Assessment of behavior, the methodology and content of psychological measurement. Columbus: Merrill, 1964.

Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. Psychometrika, 1936, 1, 53-60.

Horst, P. The prediction of personal adjustment. Social Science Research Council Bulletin 48. New York: SSRC, 1941.

Horst, P. Generalized canonical correlations and their applications to experimental data. Journal of Clinical Psychology, Monograph Supplement No. 14, 1961. (a)

Horst, P. Relations among m sets of measures. Psychometrika, 1961, 26, 129-149. (b)

Horst, P. Factor analysis of data matrices. New York: Holt, Rinehart & Winston, 1965.

Hotelling, H. The most predictable criterion. Journal of Educational Psychology, 1935, 26, 139-142.

Hotelling, H. Relations between two sets of variates. Biometrika, 1936, 28, 321-377.

Huddleston, E. M. Measurement of writing ability at the college-entrance level: objective vs. subjective testing techniques. Journal of Experimental Education, 1954, 22, 165-213.

Kaiser, H. F. Uncorrelated linear composites maximally related to a complex of correlated observations. *Educational and Psychological Measurement*, 1967, 27, 3-6.

Lawley, D. N., & Pilliner, A. E. G. Symposium: the use of essays in selection at 11+, V.- comments on "Essay marking experiments: shorter and longer essays" - by D. M. Edwards Penfold. *British Journal of Educational Psychology*, 1957, 27, 142-144.

Morrison, R. L., & Vernon, P. E. A new method of marking English compositions. *British Journal of Educational Psychology*, 1941, 11, 109-119.

Myers, A. E., McConville, C. B., & Coffman, W. E. Simplex structure in the grading of essay tests. *Educational and Psychological Measurement*, 1966, 26, 41-54.

Nyberg, V. R. A factor analytic study of essay gradings. Unpublished doctoral dissertation, University of California at Los Angeles, 1966.

Page, E. B. The imminence of grading essays by computer. *Phi Delta Kappan*, 1966, 47, 238-243.

Peel, E. A., & Armstrong, H. G. Symposium: the use of essays in selection at 11+, II.- the predictive power of the English composition in the 11+ examination. *British Journal of Educational Psychology*, 1956, 26, 163-171.

Penfold, D. M. Edwards. Symposium: the use of essays in selection at 11+, I.- essay marking experiments: shorter and longer essays. British Journal of Educational Psychology, 1956, 26, 128-136.

Pidgeon, D. A., & Yates, A. Symposium: the use of essays in selection at 11+, IV.- experimental inquiries into the use of essay-type English papers. British Journal of Educational Psychology, 1957, 27, 37-47.

Remondino, C. A factorial analysis of the evaluation of scholastic compositions in the mother tongue. British Journal of Educational Psychology, 1959, 29, 242-251.

Richardson, M. W. The combination of measures. Supplementary study D in Horst, P. (Ed.). The prediction of personal adjustment. Social Science Research Council Bulletin 48. New York: SSRC, 1941, 377-401.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Stalnaker, J. M. Weighting questions in the essay-type examination. Journal of Educational Psychology, 1938, 29, 481-490.

Stalnaker, J. M. The essay type of examination. In Lindquist, E. F. (Ed.). Educational measurement. Washington: American Council on Education, 1951, 495-530.

Starch, D., & Elliott, E. C. Reliability of the grading of high-school work in English. School Review, 1912, 20, 442-457.

Steel, J. H., & Talman, J. The marking of English compositions. London: Nisbet, 1936.

Torgerson, W. S., & Green, B. F., Jr. The factor analysis of subject-matter experts. Journal of Educational Psychology, 1952, 43, 354-363.

Travers, R. M. W. An introduction to educational research. New York: MacMillan, 1958.

Ward, J. H., Jr. Comments on "the paramorphic representation of clinical judgment". Psychological Bulletin, 1962, 59, 74-76.

Ward, J. H., Jr. Note on the independent contribution of a predictor. Psychological Reports, 1963, 12, 197-198.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

Wiseman, S. The marking of English composition in grammar school selection. British Journal of Educational Psychology, 1949, 19, 200-209.

Wiseman, S. Symposium: the use of essays in selection at 11+, III.- reliability and validity. British Journal of Educational Psychology, 1956, 26, 172-179.

Young, D. Examining essays for eleven plus classification. *British Journal of Educational Psychology*, 1962, 32, 267-274.

APPENDICES

APPENDIX A

COMPUTING ALGORITHMS

FOR

STATISTICAL PROCEDURES

APPENDIX A

COMPUTING ALGORITHMS FOR STATISTICAL PROCEDURES

The following symbols will be used in the formulae below. In order to more clearly distinguish between similar statistics, distinct symbols will be used and thus, the symbols used may differ from those used in the articles quoted.

s_i^2 = the variance of part score i

s_c^2 = the variance of the composite score

r_{ij} = the correlation between part scores i and j

c_i = the correlation between part score i and the composite

R^{ij} = the element from the i'th row and the j'th column of the inverse of the matrix of correlations between part scores

d_i = the relative importance of part score i with respect to the other part scores. Also called the desired contribution of part score i

U_i = the unique contribution of part score i, as defined by Creager and Valentine (1962), to the variance of the composite

V_i = the unique contribution of part score i, as defined by Chase (1960) and Guilford (1965), to the variance of the composite

b_i = the regression weight for part score i when all part scores are used to predict the composite in a multiple linear regression equation

W_i = the weight to be applied to variable i, as defined by Creager and Valentine (1962)

Y_i = the weight to be applied to variable i, as defined by Guilford (1965), modified as described above (pp. 48-53)

n = number of part scores making up the composite

(w) - will be used as a subscript to denote statistics for variables weighted by the Creager-Valentine weights, W_i

(y) - will be used as a subscript to denote statistics for variables weighted by the Guilford weights, Y_i

The following algorithms were used:

1.1.1 The variance of the composite was calculated by Guilford's (1965) equation 16.19 (p. 420), which may be written

$$s_c^2 = \sum_{j=1}^{n} s_j^2 + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (r_{jk} s_j s_k) \qquad (1)$$

The unique contribution of each part score to the composite variance, as defined by the Creager-Valentine method, was calculated by use of Hazewinkel's equation 4 (p. 106).

$$U_i = \frac{s_i^2}{s_c^2 \; R^{ii}} \qquad (2)$$

1.1.2 Appropriate weights were calculated for each part score by an equation derived, as shown below, from the basic equation of Creager and Valentine, (equation 3, below).

$$W_i^2 U_i s_i^2 = d_i \tag{3}$$

From (3)

$$W_i^2 = \frac{d_i}{U_i s_c^2} \tag{4}$$

and using (2)

$$W_i^2 = \frac{d_i}{s_i^2 s_c^2 \;/\; s_c^2 R^{ii}} \tag{5}$$

or

$$W_i^2 = \frac{d_i R^{ii}}{s_i^2} \tag{6}$$

and

$$W_i = \frac{\sqrt{d_i R^{ii}}}{s_i} \tag{7}$$

1.2.1 The correlation coefficient between each part score and the composite was calculated by an equation that is equivalent to Guilford's equation 16.24 (p. 427).

$$c_i = \frac{\sum_{j=1}^{n} (r_{ij} s_j)}{\sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} (r_{jk} s_j s_k)}} \tag{8}$$

The weights from a regression equation for predicting the composite from the linear sum of all part scores were calculated by

$$b_i = \sum_{j=1}^{n} (R^{ij}c_j) \qquad (9)$$

The unique contribution of each part score to the composite variance was then calculated by the method of Chase (1960) and Guilford (1965), that is from

$$V_i = b_i^2 \qquad (10)$$

1.2.2 A second set of weights for each part score was calculated by the previously described (pp. 48-53) modification of Guilford's (1965) method.

$$Y_i = \frac{\sqrt{d_i}}{s_i} \qquad (11)$$

1.3 Horst weighting system:

All calculations for finding the Horst weights were made by the use of Horst's (1965, pp. 705-708) fortran computer program. The computing algorithms are described by Horst (1961b; 1965 pp. 569-576). In order that the program could handle the amount of data involved in this part of the study it was found necessary to modify Horst's program so that only one set of weights was derived, that

yielding the maximum sum of correlations[6]. Despite this modification it was found that the program could handle only 16 of the 27 sets of marks on the computer used. A table of random numbers was used to select 16 markers from the 27.

2. Reliability of Style-content markers

The procedures used were identical to those used with the mechanics markers scores with one exception. Because there were 16 style-content part scores, as compared to 6 mechanics part scores, the computer used was able to handle the data of only 6 markers. Because the sample that could be used was so small, two random samples of the style-content markers were selected and independently analyzed by the Horst procedure. In the analysis involving the Creager-Valentine and Guilford procedures all 21 style-content markers were used.

[6] In preliminary work with the program, using three mechanics markers scores, it was found that the additional orthogonal sets of weights, that may be extracted, produced linear combinations of part scores correlating much lower with each other than did the raw scores.

APPENDIX B

WEIGHTS EXTRACTED

BY

THE VARIOUS WEIGHTING SYSTEMS

Table B1

Mechanics Markers Creager-Valentine Weights

Marker	Part Score					
No.	Spelling	Punctuation	Word Usage	Grammar	Sentence Errors	Form
1	.60	.98	1.00	1.12	1.19	1.04
2	.72	1.52	1.40	1.42	1.77	1.42
3	.61	1.02	1.22	1.07	.91	1.02
4	.66	1.06	1.24	2.76	2.98	1.07
5	.65	1.15	1.67	1.95	1.28	1.69
6	.86	1.92	2.50	1.29	2.47	1.60
7	.60	.89	1.29	.82	1.07	1.17
8	.63	1.17	1.05	1.08	.93	1.27
9	.63	.93	.93	.89	1.16	1.15
10	.64	1.07	1.34	1.28	1.64	1.61
11	.68	1.20	1.45	1.31	1.14	1.29
12	.81	2.04	1.83	1.90	1.80	1.33
13	.63	.93	.91	1.35	.86	1.32
14	.61	.78	1.02	1.53	1.10	1.28
15	.66	.93	1.31	1.17	.96	1.07
16	.63	1.18	1.47	1.47	1.16	1.59
17	.71	2.26	2.16	1.81	1.31	1.51
18	.60	1.15	1.20	1.43	1.00	1.18
19	.58	.99	1.06	.88	1.04	1.87
20	.65	1.24	1.47	1.69	1.81	1.37
21	.68	.89	1.09	1.09	1.27	1.02
22	.67	1.52	1.29	1.72	1.54	1.49
23	.64	.89	.95	.79	1.33	1.79
24	.62	1.25	1.07	1.22	1.03	1.52
25	.64	.85	1.26	.99	1.30	.89
26	.58	.79	.91	.94	.97	1.17
27	.63	1.46	1.74	1.56	1.68	1.63

Table B2

Mechanics Markers Guilford Weights

Marker	Part Score					
No.	Spelling	Punctuation	Word Usage	Grammar	Sentence Errors	Form
1	.52	.80	.82	.93	1.06	.96
2	.52	1.19	1.18	.93	1.15	1.37
3	.56	.98	1.19	1.03	.85	.94
4	.56	.96	1.01	.81	.87	1.03
5	.54	1.03	1.49	1.80	1.11	1.48
6	.54	1.38	2.25	.91	1.88	1.52
7	.53	.73	.99	.68	.89	1.07
8	.56	1.06	.99	1.04	.88	1.21
9	.55	.84	.86	.75	.96	1.09
10	.50	1.03	1.27	.99	1.13	1.56
11	.56	1.04	1.16	1.10	.89	1.12
12	.63	1.60	1.43	.88	.87	1.22
13	.51	.80	.77	1.21	.69	1.25
14	.52	.71	.83	1.39	.92	1.19
15	.56	.86	1.18	1.07	.82	1.00
16	.53	.99	1.15	1.36	.97	1.54
17	.55	1.69	1.99	1.49	1.05	1.44
18	.53	1.11	1.01	1.34	.80	1.08
19	.53	.90	.98	.74	.88	1.78
20	.59	1.19	1.29	1.23	1.31	1.26
21	.59	.81	1.00	.96	1.11	.92
22	.54	1.40	1.16	1.10	1.02	1.30
23	.51	.68	.79	.68	1.05	1.72
24	.50	1.06	.90	1.08	.87	1.46
25	.53	.71	1.04	.80	1.16	.86
26	.50	.74	.81	.80	.78	1.10
27	.58	1.27	1.40	1.30	1.22	1.49

Table B3

Mechanics Markers Horst Weights-raw scores

Marker No.	Part Score					
	Spelling	Punctuation	Word Usage	Grammar	Sentence Errors	Form
1	.07	.01	.02	.00	.01	.04
3	.07	.03	.02	.03	.03	.05
4	.07	.01	.03	-.01	.03	.07
6	.07	-.01	-.04	.03	.02	.03
7	.06	.01	.03	.01	.01	.04
8	.07	.01	.04	.04	.01	.06
11	.07	.03	.02	.01	.02	.04
13	.06	.01	.02	.02	.01	.02
14	.06	.02	.03	.01	.02	.07
15	.07	.01	.03	.02	.02	.05
20	.08	.02	.02	.00	.04	.05
21	.07	.03	.03	.01	.01	.04
22	.06	.06	.02	.02	.02	.05
23	.07	.01	.03	.01	.01	.08
25	.07	.02	.02	.00	.03	.04
27	.08	.01	.01	.02	.03	.05

Table B4

Mechanics Markers Horst Weights-z Scores

Marker	Part Score					
No.	Spelling	Punctuation	Word Usage	Grammar	Sentence Errors	Form
1	.83	.07	.14	.02	.08	.16
3	.75	.15	.08	.19	.22	.22
4	.75	.05	.15	-.05	.18	.26
6	.86	-.05	-.08	.19	.07	.10
7	.76	.08	.16	.11	.04	.15
8	.83	.05	.18	.20	.08	.18
11	.77	.13	.07	.07	.12	.12
13	.79	.08	.12	.11	.12	.06
14	.68	.18	.18	.03	.10	.22
15	.74	.07	.11	.12	.14	.19
20	.86	.07	.07	-.01	.18	.14
21	.76	.17	.15	.09	.04	.15
22	.67	.22	.10	.08	.15	.13
23	.85	.10	.17	.07	.03	.16
25	.78	.13	.08	.02	.16	.18
27	.84	.04	.04	.08	.17	.12

Table B5

Mechanics Markers Horst Weights-Spelling Omitted-z Scores

Marker No.	Part Score				
	Punctuation	Word Usage	Grammar	Sentence Errors	Form
1	.33	.19	.29	.34	.37
3	.46	.14	.32	.43	.43
4	.32	.18	-.04	.62	.38
6	.19	.01	.53	.28	.35
7	.29	.27	.22	.31	.36
8	.33	.31	.46	.41	.49
11	.30	.20	.27	.41	.31
13	.35	.37	.25	.29	.33
14	.39	.34	.16	.28	.42
15	.29	.24	.25	.38	.48
20	.31	.26	.25	.36	.36
21	.24	.16	.25	.38	.49
22	.34	.22	.11	.36	.48
23	.51	.18	.14	.51	.18
25	.27	.31	.29	.30	.36
27	.26	-.03	.32	.48	.39

Table B6

Style-content Markers Creager-Valentine Weights

Marker No.	Part Scores															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	6.4	6.7	3.5	3.3	4.2	3.6	5.5	4.7	3.9	4.8	5.5	6.6	5.8	5.8	4.8	8.6
2	3.2	4.3	3.2	3.0	3.9	3.4	6.5	4.7	4.5	4.1	5.2	6.0	4.0	5.7	3.7	4.8
3	4.3	4.8	3.5	2.8	4.0	2.9	3.9	4.0	2.8	3.9	3.7	4.3	3.5	4.9	4.2	5.1
4	6.1	5.2	5.6	4.5	4.6	5.1	5.3	5.7	4.3	5.4	5.4	5.2	4.6	5.2	6.4	6.2
5	6.5	5.4	2.9	4.5	4.3	3.1	4.7	3.9	3.2	6.1	4.9	5.7	5.0	5.3	5.9	6.9
6	8.0	6.9	5.2	3.7	4.8	5.0	5.0	6.5	4.8	4.0	9.1	7.0	5.8	6.6	6.5	4.5
7	6.6	7.4	4.4	4.2	4.3	4.1	5.7	5.3	3.8	6.5	7.9	6.7	5.6	6.9	5.0	5.4
8	6.9	7.3	2.8	3.4	4.0	3.7	4.4	5.7	3.7	3.9	2.1	5.0	4.3	4.3	4.3	6.0
9	6.4	6.0	2.9	3.6	4.1	3.2	5.9	5.0	4.3	6.1	5.6	5.9	4.5	5.9	4.4	5.3
10	4.3	4.3	4.7	3.7	3.8	4.1	5.8	5.2	4.2	4.8	5.1	5.1	4.9	6.2	6.1	5.8
11	5.6	5.4	3.9	4.0	4.5	3.0	5.3	4.4	3.3	4.0	4.8	4.8	3.8	3.6	5.3	6.8
12	5.6	4.1	2.5	3.2	4.6	3.6	4.3	4.3	3.3	2.9	4.0	3.5	4.0	4.6	4.8	4.1
13	5.6	4.7	3.9	4.0	4.4	4.1	6.5	6.2	4.0	4.9	5.7	6.5	5.8	6.6	7.1	6.4
14	9.4	9.7	3.6	3.4	4.1	3.4	4.9	4.4	3.5	6.5	4.5	5.2	5.9	5.9	8.7	7.3
15	3.9	3.1	3.5	3.2	2.8	3.2	4.6	4.6	3.7	5.0	4.9	6.2	3.3	4.5	3.6	3.5
16	6.0	7.5	3.6	4.3	3.6	3.5	4.6	4.6	3.3	3.8	3.5	5.2	3.7	5.7	4.4	6.5
17	4.9	5.7	3.3	3.2	3.9	4.1	5.7	5.6	4.2	5.3	4.5	5.9	4.3	4.4	6.1	7.1
18	5.0	4.9	5.2	3.2	3.6	4.1	6.2	5.1	4.2	5.7	4.2	5.2	5.1	5.2	5.8	6.7
19	8.8	8.6	5.2	7.8	9.0	5.7	8.8	8.9	5.8	6.9	7.9	7.2	5.0	6.5	5.8	6.4
20	7.6	5.8	4.3	4.3	4.5	4.6	6.2	6.3	5.1	6.7	4.8	9.1	7.2	7.3	7.9	6.2
21	2.9	2.8	3.0	2.8	3.2	2.7	2.8	3.9	3.1	3.8	2.6	4.0	3.0	4.1	2.6	4.3

Note: Identity of variables in Tables B6 through B9 can be found on page 69.

Table B7

Style-content Markers Guilford Weights

Marker No.	Part Scores															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	3.6	3.4	2.3	2.4	3.1	2.7	3.9	3.0	2.6	2.9	3.9	4.6	2.6	4.0	2.6	2.6
2	2.2	3.1	2.3	2.3	3.0	2.7	5.3	3.1	3.4	2.6	3.5	4.0	2.4	4.1	2.3	1.8
3	3.1	3.1	2.2	2.0	2.7	2.2	2.9	2.1	2.1	2.6	3.0	2.9	2.2	3.9	2.1	2.6
4	2.9	2.7	3.3	3.6	3.6	3.5	3.5	3.5	3.1	2.9	4.0	3.9	2.3	4.1	2.7	2.6
5	4.2	3.8	1.9	3.1	2.9	2.6	3.0	2.7	2.3	3.5	3.1	4.3	2.5	4.0	3.4	2.9
6	4.4	4.0	3.2	2.3	2.9	3.2	2.9	3.0	2.5	2.7	4.2	3.5	2.9	3.9	3.7	2.6
7	3.0	3.5	2.4	3.0	2.9	3.1	4.1	3.2	2.9	3.2	6.6	3.9	3.5	5.1	3.1	2.4
8	4.0	4.2	1.8	2.6	2.9	2.9	2.3	2.9	2.8	2.2	1.8	3.0	2.3	3.4	2.2	2.1
9	2.8	2.6	2.0	2.4	2.6	2.3	3.6	2.5	2.5	2.7	2.9	2.8	2.6	3.8	2.3	2.1
10	2.4	2.7	2.9	3.0	2.7	2.7	3.0	2.7	2.5	2.3	3.1	3.5	2.1	3.6	2.3	2.1
11	2.4	2.4	2.4	3.0	3.3	2.3	3.2	2.5	2.5	2.1	3.3	2.8	2.2	2.9	2.1	2.3
12	2.9	2.3	1.7	2.7	3.4	2.6	3.2	2.8	2.5	1.8	3.0	2.5	2.2	3.5	2.1	1.8
13	3.3	2.8	2.5	2.3	2.6	3.2	4.7	3.6	2.9	2.9	3.2	3.3	2.7	4.3	3.1	2.6
14	4.4	4.7	2.0	2.4	2.8	2.8	2.8	2.4	2.6	3.2	3.2	3.1	3.3	3.7	3.6	3.2
15	2.4	2.8	2.0	2.7	2.2	2.4	4.0	2.4	2.1	2.8	3.0	4.0	2.2	3.8	2.3	1.7
16	3.5	4.6	2.0	2.8	2.3	2.6	3.5	2.2	2.2	2.2	2.3	3.3	2.1	4.0	3.1	2.2
17	2.7	3.4	2.3	2.4	2.8	2.4	3.1	2.5	2.4	2.5	3.4	3.6	2.1	3.4	2.6	2.2
18	2.3	2.4	2.7	2.6	2.8	2.5	2.6	2.4	2.3	2.5	2.8	3.4	2.3	3.3	2.4	2.2
19	3.8	3.6	3.0	3.5	3.9	3.2	4.4	4.3	3.5	3.2	4.0	3.7	2.3	4.3	2.6	2.2
20	3.6	3.2	2.3	3.0	3.1	3.1	3.5	3.3	3.5	3.2	2.9	6.0	3.3	4.3	3.1	2.2
21	2.2	2.2	1.7	2.0	2.2	2.1	2.1	2.2	2.3	2.4	2.0	2.5	2.0	3.0	1.9	1.9

Table B8

Style-content Markers Horst Weights-Raw Scores

Marker No.	Part								Scores							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
First Random Sample of Six Markers																
1	.04	-.23	.49	-.65	-.17	.28	.14	.07	.26	.08	.10	.16	.02	-.11	-.09	-.05
5	.31	.13	.48	-.51	-.46	.27	.05	-.01	.09	-.18	.00	-.01	.05	.08	.10	-.08
7	.06	.07	.03	-.48	-.67	.26	.35	.30	.17	-.15	-.11	.49	.10	.10	-.18	.01
11	.08	.01	.61	-.76	-.51	.09	.16	.01	.24	.12	.17	.13	.16	-.13	.00	-.37
14	-.01	.42	.33	-.18	-.69	.12	.18	-.02	.18	-.23	.10	.40	.14	-.39	-.12	.03
19	-.25	.37	.21	-.67	-.89	-.04	-.16	.47	.11	.49	.30	-.29	-.04	-.03	.08	.17
Second Random Sample of Six Markers																
2	-.01	-.12	.31	.05	.34	.37	.11	.19	-.08	-.01	.41	-.07	.25	-.15	.03	.02
6	.09	-.25	.11	.20	.37	.20	.00	.27	.34	.21	.22	.06	.20	-.10	-.08	-.04
8	-.08	.27	.28	.25	.02	.19	-.17	.10	.05	.08	.09	-.19	.06	.00	.03	.17
10	.18	.07	.22	.36	-.19	.20	.24	.20	.05	-.13	.41	-.27	-.02	-.09	.16	-.03
16	.05	-.03	.18	.00	.31	.18	.21	-.52	.32	.10	.60	.22	.10	-.11	-.21	.06
20	-.17	.21	.73	.25	-.12	-.22	-.31	.06	-.11	.39	.07	.56	-.02	.40	-.16	.01

Table B9

Style-content Markers Horst Weights-z Scores

Marker No.	Part								Scores							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
First Random Sample of Six Markers																
1	.04	-.21	.47	-.61	-.12	.23	.08	.05	.22	.09	.06	.08	.03	-.06	-.13	-.07
5	.24	.11	.57	-.37	-.36	.23	.03	-.01	.09	-.16	.00	.00	.08	.05	.12	-.11
7	.06	.06	.03	-.36	-.51	.19	.19	.21	.14	-.14	-.04	.28	.11	.04	-.22	.01
11	.11	.01	.57	-.58	-.35	.09	.11	.01	.22	.18	.12	.10	.29	-.10	.00	-.64
14	-.01	.28	.37	-.16	-.54	.09	.14	-.02	.15	-.22	.07	.29	.17	-.24	-.13	.03
19	-.21	.33	.16	-.43	-.51	-.03	-.08	.24	.07	.49	.17	-.18	-.06	-.02	.12	.29
Second Random Sample of Six Markers																
2	-.01	-.13	.30	.05	.25	.31	.05	.14	-.06	-.01	.26	-.04	.41	-.08	.05	-.04
6	.06	-.20	.08	.19	.28	.14	.00	.20	.30	.25	.12	-.04	.28	-.06	-.08	-.06
8	-.07	.20	.36	.22	.01	.15	-.16	.07	.04	.11	.12	-.14	.10	.00	.05	.32
10	.24	.09	.17	.27	-.16	.17	.18	.17	.04	-.18	.29	-.17	-.04	-.06	.27	-.01
16	.05	-.02	.20	.00	.30	.15	.14	-.54	.33	.15	.59	.15	.18	-.06	-.27	.10
20	-.15	.21	.69	.19	-.08	-.15	-.20	.04	-.07	.39	.05	.21	-.02	.21	-.20	.01

APPENDIX C

STABILITY DATA

STYLE-CONTENT MARKERS

Table C1

Coefficients of Concordance

for Topics Marked at the Same Time

Marker No.	Within Marker W's		Topic No.	Between Marker W's	
	C-V wts.	G wts.		C-V wts.	G wts.
14	.767	.785	67-1	.679	.794
17	.536	.533	67-2	.715	.612
			67-3	.758	.812
			67-4	.764	.649
Medians	.65	.66		.74	.72

Table C2

Coefficients of Concordance for All Topics

Marker No.	Within Marker W's		Topic No.	Between Marker W's	
	C-V wts.	G wts.		C-V wts.	G wts.
14	.679	.746	67-1	.679	.794
17	.467	.523	67-2	.715	.612
			67-3	.758	.812
			67-4	.764	.649
			64	.794	.849
Medians	.57	.63		.76	.79

APPENDIX D

PROPORTIONS OF SHIFTS IN SCORES FOR THE DIFFERENT TRANSFORMATIONS BY MARKER

Table D1

All Mechanics Markers

Marker No.	Proportions (of total of 103 essays)						
	Adj	C-V	G	Same shift in			
				A & CV	A & G	CV & G	A, CV & G
1	.42	.41	.35	.30	.29	.33	.28
2	.19	.30	.24	.14	.13	.21	.12
3	.15	.20	.18	.12	.12	.17	.11
4	.10	.62	.17	.07	.07	.12	.05
5	.40	.48	.49	.41	.37	.41	.37
6	.11	.29	.21	.07	.05	.21	.05
7	.24	.19	.19	.11	.14	.16	.11
8	.53	.50	.50	.45	.46	.43	.41
9	.07	.14	.12	.02	.02	.10	.02
10	.10	.22	.10	.08	.07	.16	.06
11	.13	.20	.20	.09	.11	.15	.09
12	.10	.35	.18	.06	.08	.15	.05
13	.29	.26	.27	.17	.19	.22	.16
14	.13	.15	.17	.04	.08	.13	.04
15	.18	.19	.20	.15	.15	.17	.15
16	.31	.40	.42	.28	.29	.36	.27
17	.18	.25	.27	.14	.13	.24	.13
18	.04	.14	.13	.01	.01	.11	.01
19	.17	.21	.19	.12	.12	.17	.12
20	.50	.52	.49	.42	.38	.47	.37
21	.09	.15	.13	.04	.05	.10	.04
22	.12	.30	.21	.08	.08	.20	.08
23	.49	.53	.53	.41	.41	.50	.39
24	.72	.79	.75	.63	.62	.72	.59
25	.08	.20	.19	.04	.05	.16	.04
26	.50	.50	.54	.38	.43	.50	.38
27	.02	.19	.25	.02	.02	.18	.02

Table D2

All Style-content Markers

Marker No.	Proportions (of total of 103 essays)						
	Adj	C-V	G	Same shift in			
				A & CV	A & G	CV & G	A, CV & G
1	.40	.42	.41	.37	.38	.37	.36
2	.06	.10	.08	.02	.03	.07	.02
3	.11	.14	.13	.06	.09	.10	.06
4	.12	.11	.16	.06	.06	.10	.06
5	.15	.15	.17	.13	.11	.11	.10
6	.13	.14	.15	.10	.12	.13	.10
7	.19	.17	.19	.16	.16	.16	.14
8	.21	.17	.19	.16	.17	.16	.15
9	.17	.14	.14	.12	.12	.10	.10
10	.06	.07	.08	.03	.05	.04	.03
11	.07	.05	.08	.04	.01	.02	.01
12	.18	.22	.18	.12	.15	.16	.12
13	.10	.16	.15	.08	.09	.13	.08
14	.07	.17	.15	.05	.05	.15	.05
15	.17	.23	.17	.14	.15	.17	.14
16	.01	.10	.08	.00	.00	.06	.00
17	.00	.04	.05	.00	.00	.02	.00
18	.07	.12	.10	.07	.03	.03	.03
19	.08	.18	.09	.05	.05	.09	.05
20	.14	.27	.20	.11	.12	.19	.11
21	.20	.34	.26	.13	.14	.25	.13

Table D3

Mechanics Markers used for Horst Procedure

Marker No.	Proportions (of total of 103 essays)		
	Adjusted	Horst	Same Shift in Adj. and Horst
1	.32	.49	.18
3	.26	.35	.16
4	.03	.33	.02
6	.19	.31	.13
7	.12	.34	.07
8	.45	.50	.35
11	.20	.33	.13
13	.22	.36	.13
14	.05	.27	.01
15	.26	.37	.20
20	.53	.49	.31
21	.03	.36	.03
22	.05	.15	.02
23	.46	.52	.26
25	.18	.28	.11
27	.08	.34	.03

Table D4

Style-content Markers used for Horst Procedure

Marker No.	Proportions (of total of 103 essays)		
	Adjusted	Horst	Same Shift in Adj. and Horst
First Sample of Six Markers			
1	.44	.73	.17
5	.19	.50	.07
7	.18	.54	.08
11	.09	.68	.04
14	.13	.65	.04
19	.09	.40	.03
Second Sample of Six Markers			
2	.03	.37	.03
6	.16	.31	.09
8	.13	.34	.09
10	.05	.38	.03
16	.07	.47	.04
20	.14	.43	.07